Research Report

연구보고서 RE-202

글로벌 AI 경쟁력 확보를 위한 오픈소스AI 활성화 방안

Open Source AI Activation Method to Secure Global AI Competitiveness

권영환

2026.05.

SPRi 소프트웨어정책연구소

이 보고서는 2025년도 과학기술정보통신부 정보통신진흥기금을 지원받아 수행한 연구결과로 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

목 차

제1장 서 론 ····· 1

제1절 연구 배경 ····· 1

제2절 연구 내용 및 방법 ····· 3

제2장 오픈소스AI의 이해 ····· 5

제1절 오픈소스AI 생태계의 성장과 중요성 ····· 5

제2절 오픈소스AI 개념의 정립 ····· 15

제3절 요약 및 시사점 ····· 25

제3장 국내외 오픈소스AI 동향 ····· 30

제1절 현황 조사 개요 ····· 30

제2절 글로벌 오픈소스AI 기업 현황 ····· 32

제3절 국내 주요 오픈소스AI 기업 현황 ····· 45

제4절 글로벌 주요 오픈소스AI 기술(오픈소스 모델) 동향 ····· 55

제5절 요약 및 시사점 ····· 99

제4장 국내외 오픈소스AI 현황 분석 ····· 106

제1절 글로벌 오픈소스 모델 현황 분석 ····· 106

제2절 국내 오픈소스AI 인식 및 현황 설문조사 ····· 122

제3절 요약 및 시사점 ····· 148

제5장 결론: 정책적 시사점 및 정책 제언 ····· 155

그 림 목 차

[그림 1] 지속적인 오픈소스 혁신과 생태계 성장 ·····1
[그림 2] 상용 AI 서비스(gpt-o1, Claude, Gemini)와 오픈소스AI(Mistral, Llama, DeepSeek)의 비용 비교 ·····2
[그림 3] 연구 내용 구성 ·····3
[그림 4] 깃허브 AI 프로젝트 수와 관련 프로젝트들의 누적 스타 수의 증가(2011-2024) 6
[그림 5] AI 확산 및 혁신을 촉발한 오픈소스 AI 프레임워크 기술 변화 ·····7
[그림 6] AI 확산 및 혁신을 촉발한 오픈소스 AI 프레임워크 기술 변화 ·····8
[그림 7] 가트너의 생성형AI 하이퍼사이클 ·····9
[그림 8] AI 개발 기업 중 오픈소스 채택률 ·····9
[그림 9] 오픈소스AI의 장점 ·····10
[그림 10] 상용AI 서비스와 오픈 모델의 비용 비교 ·····11
[그림 11] 소버린 AI의 중요성 ·····12
[그림 12] 전세계 오픈소스 모델 누적 다운로드 수 현황(허깅페이스 기준) ·····13
[그림 13] 오픈소스AI 시스템 생태계의 예 ·····15
[그림 14] SW와 AI의 논리 구조 차이 ·····16
[그림 15] OECD의 AI 시스템 정의 ·····17
[그림 16] OSAID의 구성요소에 따른 오픈소스AI 정의의 3가지 범주 ·····19
[그림 17] OECD의 AI 시스템 생명 주기(Lifecycle) ·····21
[그림 18] 모델 개방성 프레임워크의 17개 구성 요소와 적용 라이선스 구분 ·····22
[그림 19] 모델 개방성 도구의 평가 결과의 예 ·····24
[그림 20] 모델 개방성 도구의 개방형 라이선스 정보 형태 ·····24
[그림 21] Coming for the Crown“: 글로벌 AI 모델 생태계의 지각변동 현황 ·····56
[그림 22] 라마 파생 모델 생태계 ·····61
[그림 23] 알리바바 큐웬 멀티모달 통합 아키텍처 ·····64
[그림 24] 글로벌 오픈소스 모델 현황 분석 개요 ·····106
[그림 25] 참여 기관별 모델 수 현황(전체 모델 vs 오픈소스 모델) ·····109
[그림 26] 기관 유형 현황(전체 모델 vs 오픈소스 모델) ·····110

[그림 27] 참여기관 소속 국가 현황(전체 모델 상위 20개국 vs 오픈소스 모델) ·····111
[그림 28] 연도별 모델 발표 현황(전체 모델 vs 오픈소스 모델) ·····112
[그림 29] 모델 유형 현황(전체 모델 vs 오픈소스 모델) ·····113
[그림 30] 모델 활용 분야 현황(전체 모델 vs 오픈소스 모델) ·····115
[그림 31] 유명 모델 선정 기준 현황(전체 모델 vs 오픈소스 모델) ·····116
[그림 32] 학습 데이터셋 현황(전체 모델 vs 오픈소스 모델) ·····118
[그림 33] 모델 접근성 현황 ·····119
[그림 34] 학습 코드 접근성 현황 ·····120
[그림 35] 추론 코드 접근성 현황 ·····121
[그림 36] 국내 오픈소스AI 인식 및 현황 조사 방식 ·····123
[그림 37] 주요 응답자 특성 ·····126
[그림 38] 오픈소스AI 성능 기대 ·····127
[그림 39] 오픈소스AI 노력 기대 ·····128
[그림 40] 오픈소스AI 사회적 영향 ·····129
[그림 41] 오픈소스AI 촉진 조건 ·····130
[그림 42] 오픈소스AI 저항 요인 ·····131
[그림 43] 오픈소스AI 활용 의도 ·····132
[그림 44] 오픈소스AI 실제 행동 ·····133
[그림 45] 오픈소스AI의 역량 적합도 ·····134
[그림 46] 오픈소스AI의 환경 변화 ·····135
[그림 47] 오픈소스AI의 기술 복잡성 ·····136
[그림 48] 오픈소스AI의 핵심 공개 항목 ·····137
[그림 49] 다양한 측면의 오픈소스AI 중요성 ·····138
[그림 50] 오픈소스AI 만족도 및 향후 활용 예측 ·····139
[그림 51] AI 원천 기술 도입 형태와 그 이유 ·····140
[그림 52] 오픈소스AI 적용 분야와 활용 수준 ·····141
[그림 53] 관심있는 오픈소스AI 모델 유형 ·····142
[그림 54] 오픈소스AI의 주요 장점과 단점 ·····143

[그림 55] 오픈소스AI 활용 저해 요인 및 촉진 조건 ·····144
[그림 56] 국산 오픈소스AI 중요성과 특정 플랫폼/기업 종속성 우려 ·····145
[그림 57] 오픈소스AI 전문가 육성 필요성과 전문가의 핵심 역량 ·····147

표 목 차

<표 1> 머신러닝 시스템 수정을 위해 선호되는 형태 ····· 18
<표 2> 모델 개방성 프레임워크의 3단계 개방성 수준 ····· 22
<표 3> 국내외 주요 오픈소스AI 기업 ····· 31
<표 4> 해외 주요 오픈소스AI 기업 현황 요약 ····· 44
<표 5> 국내 주요 오픈소스AI 기업 현황 요약 ····· 54
<표 6> 라마 주요 모델 ····· 58
<표 7> 메타 라마 개요 ····· 62
<표 8> 큐웬 주요 모델 ····· 63
<표 9> 알리바바 큐웬 개요 ····· 65
<표 10> 딥시크 주요 모델 ····· 67
<표 11> 딥시크 개요 ····· 69
<표 12> 미스트랄AI 주요 모델 ····· 70
<표 13> 미스트랄AI 개요 ····· 72
<표 14> 구글 주요 오픈소스 모델 ····· 73
<표 15> 구글 젬마(Gemma) 개요 ····· 75
<표 16> OpenAI 주요 오픈소스 모델 ····· 77
<표 17> OpenAI gpt-oss 개요 ····· 78
<표 18> 어니 주요 모델 ····· 80
<표 19> 바이두 어니(ERNIE) 개요 ····· 81
<표 20> 업스테이지 솔라 주요 모델 ····· 83
<표 21> 업스테이즈 솔라 개요 ····· 85
<표 22> 네이버 주요 모델 ····· 86
<표 23> 네이버 하이퍼클로버 X 개요 ····· 88
<표 24> 엑사원 주요 모델 ····· 90
<표 25> LG 엑사원 3.0 개요 ····· 91
<표 26> SK텔레콤 A.X 주요 모델 ····· 93

〈표 27〉 SK텔레콤 A.X 개요 ····· 93
〈표 28〉 NC AI의 주요 VARCO 모델 ····· 96
〈표 29〉 NC AI(구, NC소프트) VARCO 개요 ····· 97
〈표 30〉 머신러닝 시스템 수정을 위해 선호되는 형태 ····· 107
〈표 31〉 설문 조사 개요 ····· 123
〈표 32〉 설문 문항 개요 ····· 124

요 약 문

1. 제 목: 글로벌 AI 경쟁력 확보를 위한 오픈소스AI 활성화 방안

2. 연구 목적 및 필요성

소프트웨어(SW) 기술·산업 혁신을 선도해오던 오픈소스 생태계가 인공지능(AI) 기술 혁신을 선도하며 AI 산업 혁신을 촉발하고 있다. 대표적으로 오픈소스 AI 프레임워크 기술들이 언어 모델의 대형화를 선도하였고, 구글이 혁신적인 모델 구조를 오픈소스로 공개하면서 생성형 AI 시장을 태동시켰다. 그리고 메타는 OpenAI의 chatGPT에 대항하기 위해 라마를 공개하며, 빠르게 성능을 향상시며 생태계 영향력을 확대하였고 중국 기업들의 오픈소스 모델들이 공개되며 미국의 AI 주도권을 위협하고 있다.

이와 같이 오픈소스 생태계가 AI 기술·산업 혁신을 촉발하고 기술 주도권 경쟁 및 산업 영향력 확대 수단으로 활용되고 있기 때문에 AI 3강 도약을 위한 우리나라는 이러한 오픈소스 생태계 변화를 면밀히 살펴보고 우리의 나아갈 방향을 효과적으로 설정할 필요가 있다. 따라서, 본 연구는 글로벌 오픈소스AI 동향을 심층 분석하고 이를 통해 정책적 시사점과 방향을 제시하는 것을 목적으로 한다.

3. 연구의 구성

본 연구는 아래와 같이 총 5장으로 구성되어 있으며, 1장 서론으로 연구 배경과 방법에 대해 소개하고 제2장은 오픈소스AI에 대해 이해하기 위해 최근 오픈소스AI가 중요해지는 이유와 오픈소스AI 개념에 대하여 소개하고 있다.

3장부터 본격적인 본론 부분으로 국내외 오픈소스AI 동향에 대해 파악하기 위해 국내외 주요 오픈소스AI 기업을 선정하여 동향을 조사하였다. 그리고, 이들 주요 기업들의 주요 오픈소스 모델에 대한 현황을 조사하여 기술 동향을 파악하고자 하였다. 4장은 국내외 오픈소스AI 현황에 대한 실증적으로 분석하기 위해 글로벌 현황으로 EpochAI의 유명 AI 모델 데이터를 기반으로 전체 모델과 오픈소스 모델을 비교 분석하였다. 그리고 국내 오픈소스AI 인식과 현황에 대해 파악하고자 AI 개발자 대상의 설문조사를 수행하여 분석하였다.

마지막 5장은 본 보고서의 결론 부분으로 정책적 시사점을 도출하고 이를 기반으로

정책 제언을 통해 오픈소스AI 활성화를 위한 정책 방안을 제시하고 있다.

[그림 1] 연구 내용 구성

(자체 작성)

4. 연구 내용 및 결과

최근 오픈소스 기술이 SW 산업을 넘어 AI 기술 혁신과 산업 태동의 원동력이 되고 있다. 텐서플로우, 파이토치 같은 오픈소스 AI 프레임워크 기술들이 머신러닝 개발 편의성과 효율성을 제공하며 본격적인 머신러닝 중심의 인공지능 시대를 개막하는 동력이 되면서 머신러닝의 대형화를 선도하며 대형 언어 모델(LLM) 중심의 생성형AI 시장을 태동시켰다.

이러한 오픈소스 생태계의 AI 기술·산업 혁신의 근간에는 AI 분야 오픈소스 생태계의 성장이 있다. 대표 오픈소스 개발 협업 플랫폼인 깃허브의 성장과 함께 AI 관련 프로젝트가 빠르게 증가하여 약 430만개가 되었으며, 이들의 누적 스타 수도 1764만개로 매우 빠르게 증가하였다. 그리고 핵심 AI 개발 협업 플랫폼인 허깅페이스는 2025년 가입자 수가 500만명을 넘어섰고, 공개된 모델 수가 224만개가 넘어서며 빠르게 영향력을 키우고 있다.

이러한 오픈소스AI(오픈소스 생태계의 AI 기술 분야)의 빠른 성장으로 많은 개발자들이 AI 개발 과정에 오픈소스를 보편적으로 활용할 수 있게 되었다. 리눅스 재단 보고서에 의하면 AI 개발 기업 중 인프라에 오픈소스 채택 비율이 89%일 정도로 오픈소스는 AI의 기술적 기반을 제공하며 비용 절감, 위험 회피, 산업 표준(호환성 제공) 같은

긍정적 효과를 얻을 수 있게 되었다. 예를 들어 오픈소스 모델 기반 AI 서비스의 비용은 chatGPT의 1/30에 불과하며 이는 오픈소스 기술을 활용하였기 때문에 가능하게 되었다.

이와 같이 오픈소스AI가 보편적 AI 기술 인프라 역할을 수행하면서 국가적 차원의 AI 역량 강화를 위한 소버린 AI 관점에서 오픈소스AI의 중요성이 커져가고 있다. 리눅스 재단 보고서에서 국가 안보 및 경쟁력 강화를 위한 소버린 AI의 필요성에 공감하는 비율이 79%에 달하고 있으며 소버린AI의 주체로써 국가의 역할이 중요해지고 있다. 그 이유는 AI 기술 혁신이 가속화되면서 국가 경제 및 산업 활성화에 미치는 AI 영향력이 점점 커져가고 있기 때문이다. 특히 학습용 데이터 유출에 따른 개인 정보 침해, 국가 안보 위협, 기업 경쟁력 훼손이 우려되기 때문에 오픈소스AI를 활용한 자체 AI 시스템 구축 필요성이 커져가고 있기 때문이다.

오픈소스 생태계에서 AI 기술 비중이 증가하고 영향력이 확대되면서 이를 오픈소스AI라고 부르며 이 용어의 사용이 빈번해지고 있다. 따라서, 기존 SW 중심의 오픈소스 생태계와 구분하여 AI의 기술적 특징을 기반으로 오픈소스AI의 재현성과 투명성을 제고하기 위해 오픈소스AI 개념을 새로이 규정하기 위한 움직임이 있다. 대표 사례로 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 있다.

OSI의 오픈소스AI 정의는 4가지 자유(사용, 연구, 수정, 공유)가 보장된 AI 시스템으로 정의하고 있으며, 기술 범주에 따라 AI 시스템, AI 모델, AI 웨이트로 구분하며 각각에 대한 오픈소스AI 시스템, 오픈소스 모델, 오픈 웨이트로 구분하고 있다. 리눅스 재단의 모델 개방성 프레임워크는 AI 개발 생애주기 관점에서 모델 재현성을 제공하기 위한 개방성 수준을 정의하고 있다. 가장 개방적인 오픈사이언스 모델, 오픈 도구 모델, 오픈 모델로 구분하여 구분된 모델 개념간 활용 범위를 명확히 제시하고 있다.

이와 같은 오픈소스AI 태동과 새로운 개념의 규정에 따른 시사점들은 다음과 같다.

① 오픈소스AI의 중요성 : AI 기술・산업 혁신의 원동력

② 오픈소스AI의 전략적 가치 : 기술 인프라, 기업 경쟁력 요소, 소버린AI(기술 확보) 수단, 기술 경쟁 수단, 글로벌 협업 수단

③ 소버린AI를 위한 오픈소스AI의 주요 역할 : 기반 기술 제공, 외부 의존성 완화(자율성 확보), 신뢰성 확보 수단

④ 오픈소스AI 개념 정립으로 인한 기준 제시

글로벌 오픈소스AI 생태계는 AI 선도 기업들이 적극적인 기술 공개와 홍보를 통해 성장하고 있다. 특히 오픈소스AI의 핵심이라고 할 수 있는 오픈소스 모델은 선진 AI 기업들이 주도하고 있다. 따라서, 오픈소스AI 생태계를 선도하는 국내외 기업과 이들 기업

들의 오픈소스 모델을 중심으로 글로벌 오픈소스AI 동향을 조사하였다.

대상 기업으로 글로벌 오픈소스AI 생태계를 선도하는 해외 기업 8개(메타, 구글, OpenAI, EleutherAI, 알리바바, 바이두, 딥시크, 미스트랄AI)를 선정하였다. 조사 결과, 기업들은 AI 기술을 기반으로 자체 플랫폼 경쟁력 강화, 상용AI 서비스 제공, AI 솔루션 공급을 하고 있다. 그리고 이들 기업들의 오픈소스AI는 단순한 기술 공개를 넘어 특정 기업 견제(종속성 회피), 생태계 영향력 확보 및 확장, 기술력 입증, 개방형 협업, 글로벌 진출, 기술 주권 확보 등 다양한 목적 달성을 위한 전략적 수단으로 볼 수 있다.

〈표 1〉 해외 주요 오픈소스AI 기업 현황 요약

구분	주요 오픈소스 AI	법인 유형	AI 관련 주요 사업	오픈소스 전략	주요 파트너십
메타	Llama 3.1 (405B) Llama4	상장사	SNS 서비스 AI 전환, 광고 최적화, AI 서비스	OpenAI 견제(종속성 회피) 및 자체 플랫폼 경쟁력 강화	Microsoft Azure
구글	BERT (11만+ 인용) Gemma2(27B)	상장사	검색, 광고, 클라우드, 온디바이스 AI	자체 플랫폼 경쟁력 강화 및 생태계 확장(온디바이스)	AWS, GCP AndroidOEM 클라우드고객
OpenAI	GPT-2, gpt-oss-120B	비영리→영리 전환	AI 서비스, 기술 공급	영향력 확보(기술력 입증)	Microsoft (독점) $13B투자
EleutherAI	GPT-Neo (2.7B) GPT-J-6B Pythia	비영리	오픈소스 협업 (커뮤니티)	개방형 협업, 학술공헌	CoreWeave StabilityAI후원
딥시크	DeepSeek-V3 (671B) DeepSeek-R1(추론 모델)	스타트업	AI 서비스, 기술 공급	영향력 확보(기술력 입증)	API 플랫폼 (독립 운영)
알리바바	Qwen2.5 (0.5B-72B) Qwen3 (235B, MoE)	상장사	클라우드 AI, 전자상거래	플랫폼 경쟁력 강화 및 글로벌 영향력 확보	Alibaba Cloud 전자상거래통합
바이두	Ernie 4.5 ErnieSpeed	상장사	검색, 자율주행, 클라우드	시장 경쟁력	Baidu Cloud Apollo자율주행
미스트랄 AI	Mistral 7B Mixtral8x7B Codestral	스타트업	기술 공급	유럽 AI 주권 미국/중국견제	Microsoft Azure AWS 유럽기업

(자체 작성)

국내 기업으로는 오픈소스 AI를 선도하는 기업 5개(네이버, LG AI Research, SK텔레콤, 업스테이지, 엔씨 AI)를 선정하였다. 조사 결과, 국내 기업들은 보편적으로 기술력 입증을 위한 모델 공개 전략을 추진하고 있는 것으로 보인다. 아직은 국내 기업들의

AI 기술력의 인지도 부족과 생태계 영향력 부족으로 인한 것으로 기술 공개를 통해 전략적 입지 마련을 위해서로 생각된다. 향후에 국내 오픈소스 모델이 글로벌 오픈소스 모델 대비 성능이 견줄 수 있다면 해외 진출을 추진할 수 있다고 보며, 한국어 처리 능력에서 우월성이 증명된다면 국내 시장의 해외 오픈소스 모델의 점유율 하락을 기대할 수 있을 것으로 생각된다.

〈표 2〉 국내 주요 오픈소스AI 기업 현황 요약

구분	주요 오픈소스 AI	법인 유형	AI 관련 주요 사업	오픈소스 전략 목적	주요 파트너십
LG AI	EXAONE 3.0 (7.8B) EXAONE Universe	대기업 연구 조직	제조 산업 특화 (가전, 화학, 신약 등) ChatEXAONE (사내용),	기술력 입증	LG 화학/에너지 솔루션
SK텔레콤	A.X (에이닷 엑스)	대기업	스마트폰 AI 서비스(AI 비서, 통화 번역) 고객서비스 향상 및 효율화	글로벌 통신사 LLM 표준화 추구	OpenAI, Anthropic, 도이치 텔레콤, 싱텔
네이버	HyperCLOVA X HyperCLOVA X SEED	대기업	검색(Cue:) 강화 및 쇼핑 추천, 클라우드 서비스	전문가 생태계 활성화, 기술력 입증	인텔 (AI 반도체), 사우디 아람코
업스테이지	Solar Pro (10.7B/22B) Solar Mini	스타트업	문서 특화, AI 서비스 및 사설 LLM 구축	기술력 입증 -> 글로벌 인지도 향상	AWS, 삼성생명, 신한은행
엔씨AI	VARCO LLM VARCO Art/Text	대기업 자회사	게임 특화, AI NPC, QA 자동화	기술력 입증	Unity/Unreal 엔진 (3D 플러그인)

(자체 작성)

국내외 오픈소스AI 기업 동향을 통해 도출한 시사점들은 다음과 같다.

① 오픈소스AI 전략의 실질적 목표: 기술・산업 영향력 확대

② 중국 기업의 차별화 오픈소스 전략: 적극적 공개(혁신 기술 + 오픈소스 라이선스)

③ 오픈소스AI 영향력 확대 방안: 빠른 기술 혁신과 잦은 반복 출시로 개발자 락인

④ 다양한 오픈소스AI 공개 방안 : 완전 공개, 제한적 공개, 폐쇄 전략

⑤ 오픈소스AI 수익화 모델 : 간접 수익화 - 자체 제품/서비스 혁신 및 로스 리더(Loss Leader) 전략

⑥ 오픈소스AI의 주요 수익원 : B2B 시장(엔터프라이즈 고객)

4장은 국내외 오픈소스AI 생태계에 대한 실증적 분석을 위해 EpochAI의 유명 AI 모델 데이터를 기반으로 글로벌 AI 생태계에서 오픈소스 모델의 영향력과 현황을 분석하였고 국내 현황으로 371명의 AI 개발자들의 오픈소스AI 인식 및 현황을 파악하기 위해 설문 조사를 수행하여 분석하였다.

EpochAI의 유명 AI 모델 데이터는 1950년부터 2025년 6월 초까지 발표된 AI 모델 중 4가지 조건(① 공인 벤치마크의 최첨단 개선, ② 인용 수 1000개 이상, ③ 기술 발전의 중요성, ④ 중요한 활용)중 하나 이상을 충족한 964개의 AI 모델 정보를 제공하고 있다. 이 데이터를 기반으로 OSI의 오픈소스AI 정의를 기반으로 주요 공개 항목(데이터, 모델 구조, 웨이트, 학습 & 추론 코드 등)을 중심으로 오픈소스 모델 269개를 분류하여 참여 기관, 참여 기관 국가, 조직 분류, 발표 일시, 유형, 활용 분야, 선정 기준 등을 분석하였다.

참여 기관에서 가장 많은 유명 모델 개발에 참여한 기관은 구글로 타 기관 대비 압도적이었으며 오픈소스 모델 공개는 메타(페이스북 포함시)가 45개로 1위를 차지하였다. 그리고, 기관 유형은 산업계가 참여한 AI 모델이 626개 이었으며, 오픈소스 모델은 208개로 기업들이 AI 기술 혁신을 주도하고 있었다.

연도별 현황을 보면 AI 모델은 2013년부터 지속적으로 증가하고 있으며, 오픈소스 모델은 2015년에 처음 등장하여 2018년부터 지속 증가하는 추세이다. 그리고, 2019년 이후 유명 모델의 51.9%는 오픈소스 모델일 정도로 오픈소스 생태계 영향력이 최근 빠르게 증가하고 있다.

모델별 참여 국가 현황을 보면 미국이 AI 모델과 오픈소스 모델 모두 1위였고 이어서 중국이 2위 었으며, 우리나라는 오픈소스 모델 분야에서 양적으로 7위에 위치하고 있었다. 이러한 데이터는 최근 AI 생태계에서 중국 기업의 부상을 데이터로 입증하는 자료이며 허깅페이스 자료에서도 중국 모델들이 전세계적으로 주목을 받고 있음을 보여주고 있다.

데이터 정보 공개, 모델 접근성, 학습 코드 접근성, 추론 코드 접근성의 분석 결과를 보면 유명 모델의 약 절반인 48.3%(458개) 모델들이 학습 데이터 정보를 공개하고 있으며 학습 데이터 종류는 276개로 다양한 데이터들이 활용되고 있음을 알 수 있다. 이에 반해 모델 접근성, 학습 코드 접근성, 추론 코드 접근성을 공개한 모델은 각각 269개, 237개, 225개로 상대적으로 적음을 알 수 있다.

국내 오픈소스AI 인식 및 활용 현황 조사는 선행 문헌들을 분석하여 총 60개의 설문 문항을 개발하여 모바일 온라인 설문 조사를 통해 총 371명의 AI 개발자로부터 응답 결과를 수집하였다. 주요 구성은 Part 1은 성능 기대, 노력 기대, 사회적 영향, 촉진 조건, 저항 요인, 활용 의사, 사용 행동, 기술 적합성, 환경 역동성, 제품 난이도에 대해 40개의 리커드 척도 문항으로 구성하였다. Part2는 오픈소스AI 관련 개발자들의 직접적

인 인식과 상세 현황을 위한 설문으로 핵심 공개 항목, 중요성, 만족도 및 향후 예측, 국산 중요성, AI 도입 유형, 적용 분야 및 활용 수준, 모델 유형, 장단점, 저해 요인 및 촉진 조건, 플랫폼/기업 종속성, 전문가 필요성 및 핵심 역량에 관한 20개의 문항으로 구성하였다.

Part 1의 오픈소스AI 활용 동기 측면에서 성능 기대에서 업무 생산성과 빠른 업무 수행에 상대적으로 많은 응답자들이 동의하였으며, 노력 기대에서 필요 정보 획득에 가장 많은 응답자들이 선택하였으며, 사회적 영향에서 의사결정자와 고위 경영진이 상대적 으로 많은 응답자들이 동의하였고, 촉진 조건에서 필요 지식이 가장 많이 응답되었다. 저해 요인으로는 저작권 침해에 가장 많은 응답자들이 동의하였다.

AI 개발자들의 오픈소스AI 활용 현황을 보면 활용 의사에서 보면 주변 추천, 자주 사용, 지속 사용, 방법 탐색 등 다양한 측면에서 활용을 확산할 의사를 가지고 있음을 확인하였다. 그리고 실질적 행동으로 기술 문서 탐색, 새로운 정보 획득, 학습 자료 활용 등을 통해 오픈소스AI 역량을 강화하려는 응답 비중이 50%를 넘었다.

Part 2에서 개발자들은 오픈소스AI의 핵심 공개 항목으로 모델, 데이터 정보, 학습 데이터, 기술 문서를 선택하였다. 이는 공개된 정보를 기반으로 활용과 성능 향상을 위한 핵심 정보들도 판단된다.

다양한 측면의 오픈소스AI 중요성에 대해 개인적인 업무상 중요성 외에 국가 AI 기술력 강화, 기업 AI 경쟁력 강화, AI 기술 확산 측면에 오픈소스AI 중요하다는 응답에 90% 안팎으로 응답할 정도로 중요하게 인식되고 있었다. 그리고 오픈소스AI에 만족하는 개발자 비율이 54.4%로 과반이 넘었고, 향후 활용이 확대될 것으로 예상하는 개발자 비율이 89%이었다. 실제로 AI 원천 기술 도입 형태에서 오픈소스AI를 활용하는 비율은 68.5%이었다.

적용 분야와 활용 수준을 보면 사내 업무 및 서비스 적용과 기존 제품/서비스 개선이 각각 60.4%와 59.0%로 새로운 제품/서비스 개발에 활용하는 비율보다 높았으며, 활용 수준도 제품 적용 보다는 기술 탐색 및 연구개발 진행 중이 전체의 72.3%로 아직까지는 제품 적용 단계는 소수인 것으로 파악되었다.

그리고 오픈소스AI 기술의 장단점으로 빠른 기술 발전 속도, 라이선스 비용 절감, 및 기술 내재화 가능성에 많은 응답자들이 선택하였으며, 단점으로는 기술 완성도 부족, 책임 주체의 불명확성, 부족한 문서화 등이 선택되면서 기술 역량 확보가 오픈소스AI 확산의 주요 요소로 판단된다.

Part1의 오픈소스AI 저항 요인과 촉진 조건을 상세히 묻는 설문에서 주요 저해 요인으로는 조직의 오픈소스AI 전략(비전) 부재, 인력 부족, 자체 인프라 부족, 조직의 정책/거버넌스 부재 순으로 응답되며 개인적 오픈소스AI 활용 의사에 비해 환경적 요인이 부족한 것으로 판단된다. 그리고 주요 촉진 조건으로는 연구개발 인프라, 전문가 양성

및 발굴, 기업/기관 거버넌스 확립, 오픈소스AI 활용 기업 발굴 등이 응답되며 오픈소스AI 활용 기반 조성 및 인력 양성 필요성이 제기되었다.

기술 종속성 관점에서 국산 오픈소스AI의 중요성과 특정 플랫폼/기업 종속성에 대한 설문 결과를 보면 국산 오픈소스AI 기술이 중요하다고 응답한 비율이 해외 오픈소스AI가 중요하다고 응답한 비율보다 10.2% 높았으나, 둘 다 중요하다고 응답한 비율이 절반이 넘는 50.9%임을 고려하면 개발자들은 국산 혹은 외산 구분 보다는 우수한 오픈소스AI에 관심이 많다고 추정할 수 있다. 다만, 특정 플랫폼 및 기업 종속성에 대한 우려에 대한 응답은 50.7%로 이에 대한 우려가 있음을 확인하였다.

마지막으로 오픈소스AI 확산의 주요 저해 요소인 인력(개발자, 기여자) 부족이자 주요 촉진 조건인 전문가 양성을 위한 오픈소스AI 전문가 육성 필요성에 중요하다고 응답한 비율은 88.1%로 많은 개발자들이 동의하고 있었다. 그리고, 전문가의 주요 역량으로 AI 모델 개발, 오픈소스 프레임워크 활용, 데이터 처리 및 품질 관리와 데이터셋 구축 역량, 오픈소스 기여 및 협업 역량 순이었다.

이러한 국내외 오픈소스AI 현황 분석을 통해 다음과 같은 시사점들을 도출하였다.

① 빠르게 성장하는 오픈소스AI 생태계
② 기업/국가 경쟁력 강화를 위해 중요한 오픈소스AI
③ 기업 인식 개선 필요
④ 오픈소스AI 활용을 위한 전제조건 : 기술 역량 확보
⑤ 오픈소스AI 인재 양성 필요
⑥ 자주적 AI 역량 확보가 필요

최근 오픈소스AI가 빠르게 성장하면서 AI 기술 혁신 및 산업 혁신의 원동력이 되면서 AI 3강 도약을 위한 오픈소스AI 정책의 필요성이 커져가고 있다. 따라서 마지막 5장에서 본 연구 내용을 기반으로 다음과 같은 정책적 시사점과 정책을 제언하고자 한다.

① 선진 기술의 빠른 수용을 위한 오픈소스AI 활용 확산 지원
② AI 기술력 강화를 위한 선진 오픈소스AI 기술 내재화 추진
③ 기업의 오픈소스AI 인식 개선 필요
④ 오픈소스AI 전문가 양성 및 교육 확대 필요
⑤ 오픈소스AI 전문 컨설팅 제공 필요
⑥ 오픈소스AI 전문기업 육성

5. 정책적 활용 내용

AI의 중요성이 급부상하면서 우리나라는 최근 AI 3강 도약을 위한 투자 및 정책적 지원이 집중되고 있다. 최근 글로벌 AI 동향을 보면 AI 선도 기업을 중심으로 오픈소스 모델을 공개하며 기술·산업 주도권 확보 경쟁이 치열해지면서 오픈소스AI 생태계가 중요해지고 있다. 특히 오픈소스AI는 글로벌 기업 중심의 AI 대외 의존도를 완화할 수 있는 대안을 제시하고 있다. 이러한 현실에서 본 보고서는 국가 AI 경쟁력 강화를 위한 오픈소스AI 정책 수립을 지원을 위한 연구 목적으로 작성되었다.

따라서, 본 보고서는 AI 3강 도약을 위한 오픈소스AI 정책을 위한 당위성 제공과 정책 방향 수립를 위한 핵심 근거 자료를 제공할 수 있으며, 국가 AI 경쟁력 확보를 위한 오픈소스AI 기반 기술 확보 방안, 오픈소스AI 기업 육성, 오픈소스 생태계 활성화 정책 수립의 근거 자료로 활용될 수 있다.

6. 기대효과

오픈소스AI는 국가 AI 기술 경쟁력 강화, 기업 AI 경쟁력 강화, AI 기술 확산 등 다양한 관점에서 중요하다. 특히 글로벌 기업 중심의 AI 기술 혁신에 의해 해외 의존도가 심화되고 있다. 오픈소스AI는 대외 의존도를 완화할 수 있는 가장 효율적인 대안으로 자주적인 AI 역량 확보를 통한 산업 경쟁력 강화에 큰 기여를 할 수 있다. 따라서, 본 연구는 오픈소스AI 정책 수립을 위한 기초 자료 제공과 정책 방안 제시를 통해 국가 AI 경쟁력 제고를 위한 AI 기술력 확보, AI 기업 경쟁력 강화를 통한 AI 산업 활성화 등 다양한 측면에서 긍정적 효과가 기대된다.

제1장 서 론

제1절 연구 배경

과거 오픈소스(Open Source) 생태계에서 개발된 기술들이 소프트웨어(SW) 기술 혁신과 산업 혁신을 넘어 최근에는 인공지능(AI) 기술 혁신과 산업 혁신을 선도하고 있다. 초기 오픈소스 생태계는 파편화된 SW 생태계(유닉스 운영체제)의 호환성 향상을 위한 개발자 중심의 생태계로 1980년대에 시작하였다. 1990년부터 리눅스 커널(운영체제), 데이터베이스(DB), 웹 서버 분야로 확장하며 상용SW의 대안으로 주목을 받았다. 2000년대가 되면서 DB 기술이 빅데이터 기술로 고도화되고, 서버 기술들이 클라우드 기술로 진화하면서 플랫폼 SW 생태계를 촉발하였으며, 다양한 산업 분야로 확산되며 SW 융합의 확산을 주도하였다.

최근 오픈소스 생태계는 AI 기술 혁신을 선도하고, 산업 생태계 변화를 촉발하면서 다시 한번 크게 주목받고 있다. 2025년 1월 혁신적인 딥시크-R1의 공개는 낮은 비용의 학습이 가능하다는 것을 증명하며 대형 언어 모델(LLM)개발이 빅테크 기업의 전유물이 아님을 입증하였다. 그리고 메타 라마(LLama)의 제한적 공개(월간 7억 이상 사용 제한)를 넘어 자유로운 사용이 가능한 MIT 라이선스로 공개하였다. 이와 같이 AI 기술의 오픈소스 공개는 AI 기술 혁신 가속화와 기술 확산을 통해 산업 생태계를 변화시킨다는데 큰 의미가 있다.

[그림 1] 지속적인 오픈소스 혁신과 생태계 성장

시대	1980년대 (태동)	2000년 전후 (성장)	2010년 전후 (고도화)	2020년 전후 (확산)	2025년 현재 (확장)
SW 규모	**시스템SW의 구성요소**	**시스템SW**	**플랫폼SW**	**산업 맞춤형 플랫폼**	AI = SW + **데이터**
주요 오픈소스	GNU 프로젝트 (X.11 - 윈도우 환경, Gdb - 디버거, Gcc - 컴파일러, Bash – SW 제어 환경. 등)	리눅스 커널 (운영체제, 1991), MySQL (DB, 1995), 아파치 HTTP 서버 (웹 서버, 1995), 모질라 (브라우저, 1998)	아파치 하둡 (빅데이터, 2006), 안드로이드 (모바일, 2008), 오픈스택 (클라우드, 2010), Automotive Grade Linux (커넥티드카, 2012), 텐서플로우 (인공지능, 2015)	Fintech Open Source (2017, 금융), Academy Software (2018, 미디어), LF Networking (2018, 통신/5G), LF Energy (2018, 에너지), OS-Climate (2020, 기후 온난화), Green Software (2021, 친환경 SW)	메타 (라마), 미스트랄AI, 딥시크, 알리바바 (큐웬), 바이두 (어니), OpenAI (gpt-oss)
목적/역할	SW 개발 호환성 제공	상용SW 대안	SW/IT 신기술 개발 협력	산업 디지털 혁신 및 지구 문제 대응	AI 혁신
산업 분야	SW 개발자	SW 산업	ICT 산업	전 산업	신산업(AI)
IT 생태계 파급 효과	SW개발 저변 확대 (C, 윈도우 환경, 등)	인터넷 산업 활성화	플랫폼 비즈니스 확산	산업 디지털 전환	AI 전환
핵심 개념	**자유 소프트웨어**	**오픈소스(SW)**	**엔터프라이즈 오픈소스**	….	**오픈소스AI**

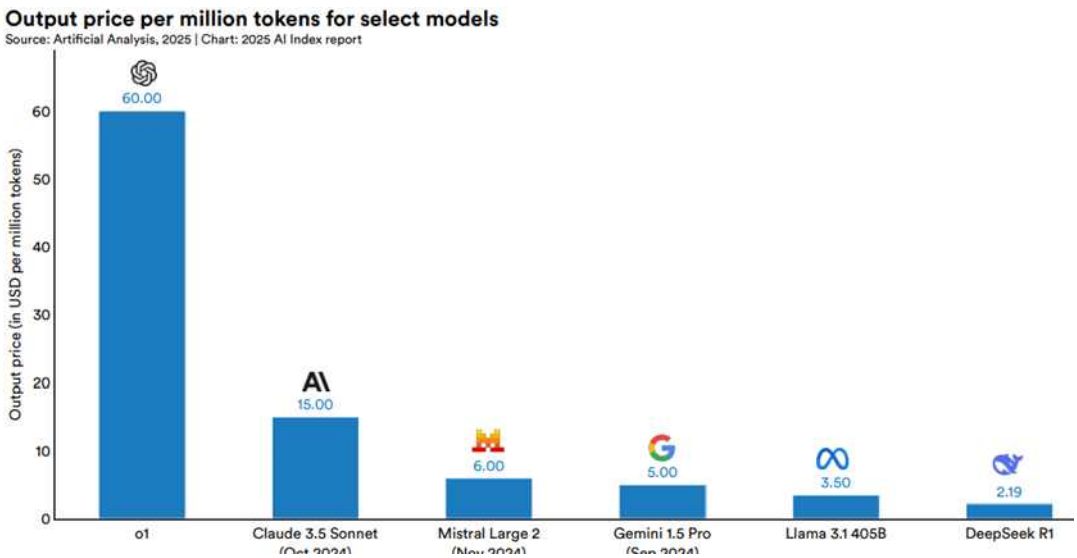
(자체 작성)

따라서, AI 3강 도약을 국가적을 추진하고 있는 우리나라는 오픈소스 생태계의 AI 기술 혁신과 이를 통해 촉발되는 산업 생태계 변화에 주목할 필요가 있다. 특히 미국과 중국이라는 초거대 국가들의 기술 주도권 경쟁과 선업 경쟁이 가속화되는 현실속에서 자주적인 AI 역량을 확보하고 신 AI산업 육성 및 국가 주력 산업의 경쟁력 강화는 미래 국가 경쟁력의 핵심이기 때문에 오픈소스 생태계의 AI 동향 분석은 매우 중요하다.

특히 최근 공개되는 오픈소스 AI 기술 중에 오픈소스 모델의 성능은 최신 상용 AI 서비스의 성능에 근접하고 있으며, 우월한 낮은 가격(chatGPT 대비 약 1/30 수준)으로 주목받고 있다. 그러므로, 기술이 공개된 오픈소스AI 기술의 활용은 국내적으로 외국 기업 의존도를 완화할 수 있는 효율적인 수단이 될 수 있으며, 국제적으로는 우수한 오픈소스AI 기술을 내재화하여 해외 진출 교두보를 구축할 수 있는 기반이 될 수 있기 때문에 AI 3강 도약의 실현 가능성과 성과 확대를 위한 중요한 전략적 자산이 될 수 있다.

따라서, 본 연구는 최근 오픈소스 생태계에서 벌어지는 AI 기술 혁신과 변화를 면밀히 분석하여 AI 3강 도약 및 자주적인 AI 역량 확보를 위한 정책적 시사점을 제시하고 이를 통한 실질적 정책 제언을 하고자 한다.

[그림 2] 상용 AI 서비스(gpt-o1, Claude, Gemini)와 오픈소스AI(Mistral, Llama, DeepSeek)의 비용 비교



출처) 스탠포드 대학 HAI, AI Index 2025. 2025.04.

제2절 연구 내용 및 방법

본 연구는 서론(1장)에서 본 연구의 배경과 필요성을 제시하고 2장부터 4장까지는 본 연구의 핵심인 글로벌 오픈소스 생태계의 AI 동향과 국내 현황에 대해 소개한다. 그리고 마지막을 결론(5장)에서 본 연구 내용을 요약하고 후속 연구 방향을 제시하며 마무리한다.

[그림 3] 연구 내용 구성

(자체 작성)

2장은 오픈소스AI의 부상과 개념 정립에 대해 소개하는 장으로 최근 오픈소스AI가 주목받는 이유와 오픈소스AI에 대한 다양한 개념들을 소개하고 있다. 오픈소스AI가 최근 주목받는 이유는 첫 번째로 오픈소스AI가 SW를 넘어 AI 기술 혁신과 확산을 선도하고 있기 때문이다. 실제로 theano, 텐서플로우, 파이토치 등이 머신 러닝 시대를 개박하고 모델 대형화를 지원하며 생성형 AI 시대를 선도하고 있다.

이렇게 오픈소스AI의 중요성이 커짐에 따라 모호한 오픈소스AI 개념 정립을 위한 노력들이 자생적으로 발생하였다. 대표적으로 OSI의 오픈소스AI 정의와 리눅스재단의 모델 개방성 프레임워크가 있다. 본 연구에서는 이들 개념을 상세히 분석하여 소개하고 있다.

3장은 글로벌 오픈소스AI 동향을 파악하기 위해 국내외 오픈소스AI 기업과 해당 기업들의 주요 오픈소스AI 기술(오픈소스 모델 중심)의 동향을 조사하였다. 해외 기업으

로는 오픈소스AI 생태계 영향력을 기반으로 메타, 구글, OpenAI, Eleuther AI, 중국의 알리바바, 바이두, 딥시크, 유럽의 미스트랄 AI를 선정하여 조사하였고, 국내 기업으로는 독자 AI 파운데이션 모델 개발 사업을 추진하고 있는 네이버, LG AI, SK 텔레콤, 업스테이지, 엔씨 AI를 선정하여 조사하였다. 주요 조사 항목은 오픈소스AI 관련 일반 현황과 사업화 현황을 중심으로 하였다.

그리고 기술 동향을 파악하기 위해 이들 기업들의 주요 모델에 대한 조사를 수행하였다. 대상 기술로는 메타 라마, 구글 젬마, OpenAI gpt-oss, 알리바바 큐웬, 바이두 어니, 딥시크, 미스트랄, LG 엑사원, 네이버 하이퍼클로버 X, 업스테이지 솔라, SK 텔레콤 A.X, 엔씨AI VARCO를 대상으로 하였다. 이들 오픈소스 모델 기술에 대해 주요 공개 시점, 라이선스, 허깅페이스 현황 및 개발 지표(다운로드 수, Likes 수, 팔로워 수, 커뮤니티 수 등), 주요 활용 사례 들을 조사하였다.

4장은 국내외 오픈소스AI 현황에 대한 실증 분석 차원에서 글로벌 현황 분석으로 EpochAI의 Notable AI 모델 데이터를 분석하였고, 국내 현황 분석으로 오픈소스AI 인식 및 현황 설문 조사 결과를 분석하였다.

EpochAI는 AI 생태계 전반에 대해 연구하는 미국의 비영리 기관으로 1950년 이후의 유명 AI 모델 정보를 축적하여 제공하고 있다. 이 자료를 기반으로 전체 모델과 오픈소스 모델의 다양한 현황을 분석하였다. 주요 분석 항목으로는 참여 기관별 개발 참여 모델 현황, 참여기간 국가별 모델 수, 연도별 발표 모델 수, 모델 유형별 현황, 모델 활용 분야별 현황, 유명 모델 선정 기준 등을 분석하였다.

국내 오픈소스AI 인식과 현황을 파악하기 위해 선행 문헌들을 분석하여 60개 문항을 도출하여 국내 AI 개발자 371명을 대상으로 설문 조사를 수행하여 해당 결과를 분석하였다. 주요 설문 내용은 오픈소스AI 활용에 미치는 다양한 요인들을 분석하기 위한 Part 1에서 성능 기대, 노력 기대, 사회적 영향, 촉진 조건, 저항 요인, 활용 의사, 사용 행동, 기술 적합성, 환경 역동성, 제품 난이도 등을 조사하였다. 그리고 상세 인식과 현황을 파악을 위한 Part 2에서 핵심 공개 항목, 중요성, 만족도 및 향후 예측, 구산 중요성, AI 도입 유형 및 이유, 적용 분야 및 활용 수준, 모델 유형, 장점 및 단점, 저해 요인 및 촉진 조건, 플랫폼/기업 종속성, 전문가 필요성 및 핵심 역량 들을 조사하였다.

그리고, 마지막 5장에서는 결론으로 정책적 시사점과 정책 제언을 통해 보고서를 마무리 하였다.

제2장 오픈소스AI의 이해

제1절 오픈소스AI 생태계의 성장과 중요성

오픈소스 생태계는 과거 SW 기술·산업 혁신을 선도하였으며 현재는 인공지능 기술 혁신과 산업 혁신의 중심이 되고 있다. 1980년대 태동되었던 초기 오픈소스 생태계는 SW 호환성 향상 및 협업 활성화를 위한 개발자 중심 생태계로 출발하여 1990년대에 들어서 독점 SW 기업에 대항할 수 있는 운영체제, DB, 웹 분야의 기술적 기반을 제공하였다.

대표적으로 리눅스 커널, MySQL 같은 오픈소스 DB, 아파치 웹 서버 기술들이 있으며, 이들은 상용SW의 대안으로 주목받으며 인터넷 산업 활성화에 크게 기여하였다. 실제로 전세계 서버 시장(450억 달러 규모)에서 리눅스와 오픈소스 웹서버 비중은 모두 70% 이상[1]이며, DB 분야의 오픈소스DB 비중은 약 50%[2]일 정도로 SW 산업에서 영향력을 확보하고 있다.

2000년대에 등장한 아파치 하둡(2006). 안드로이드(2008), 오픈스택(2010), 텐서플로우(2015) 등의 오픈소스 플랫폼 기술들은 기존 오픈소스 기반 운영체제, DB, 웹 서버 기술들이 상호 융합하면서 플랫폼 비즈니스 활성화를 주도하였다. 특히 안드로이드는 모바일 플랫폼 시장을 석권하여 모바일SW 시장의 성공을 가져왔고, 아파치 하둡, 오픈스택, 텐서 플로우들은 각가 빅데이터 플랫폼, 클라우드 플랫폼, 인공지능 플랫폼의 대표 주자가 되면서 새로운 플랫폼 기반 SW시장의 급성장의 기폭제가 되었다.

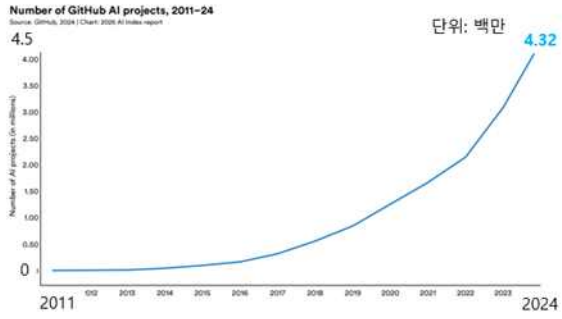
오픈소스 기술이 비용 절감, 개발 속도 향상 등의 SW 경제성의 중요한 요소로 부상하면서 2015년 이후부터 타 산업 분야에서도 우수한 오픈소스 기술을 활용한 산업 맞춤형 플랫폼 확보를 위한 노력이 확산되고 있다. 대표적으로 금융 분야의 FINOS(FINtech Open Source) 재단, 미디어 분야의 Academy Software 재단, 통신 분야의 LF Networking 재단, O-RAN 얼라이언스, 에너지 분야의 LF 에너지 재단, 자동차 분야의 AGL(Automotive Grade Linux)와 아폴로(Apollo) 프로젝트 등이 있다. 이들은 해당 산업 기업들과 IT 기업들이 오픈소스 개발 협업을 통해 산업 맞춤형 플랫폼 기술 개발과 시장 창출에 힘쓰고 있다.

---

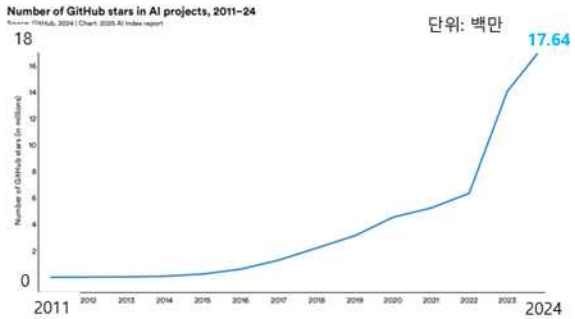
1) Gitnux, Report 2025 - Server Statistics, 2025.04.29.
2) DB Engines, Popularity of open source DBMS versus commercial DBMS, 2025.11.20.

2020년대에 들어서는 기존 오픈소스 기반 빅데이터, 클라우드, 인공지능 기술이 상호 융합되며 AI 기술 혁신과 신산업 확산의 중심이 되고 있다. 대표 오픈소스 개발 협업 플랫폼인 깃허브와 허깅페이스에서 AI 분야의 오픈소스개발 협업이 빠르게 증가하며 기술 혁신 및 확산을 가속화 시키고 있다. 2025년 깃허브 개발자 수는 3600만명의 개발자가 증가하여 누적 가입자 수는 1억 8천만명으로 성장하였다. 그리고 2024년 깃허브는 4,320만건의 풀리퀘스트(기여, +23% 증가), 약 10억건의 커밋(코드 변경 승인, +25.1% 증가), 6.3억개의 프로젝트로 증가하였는데, 그 중심에는 430만개 이상의 AI 프로젝트들(vllm-고처리량 LLM 추론 엔전, ollama-로컬 모델 실행 및 관리 도구, transfomers-모델 구조, ragflow - RAG(Retrieval-Augmented Generation) 엔진)이 있으며, 누적 스타 수도 1764만개 이상으로 빠르게 증가하였다[3)4)].

[그림 4] 깃허브 AI 프로젝트 수와 관련 프로젝트들의 누적 스타 수의 증가(2011-2024)



(a) 깃허브의 AI 프로젝트 수



(b) 깃허브의 AI 프로젝트 누적 스타 수

출처) GitHub Octoverse 2025.

세계 최대의 오픈소스 모델 개발 협업 플랫폼인 허깅페이스도 매우 빠르게 활성화되고 있다. 허깅페이스는 2016년 설립된 AI 스타트업으로 깃허브에서 개발되는 transformers를 기반으로 최신 AI 모델의 효율적 개발, 데이터셋 공유, 강력한 엔비디아 추론 기능을 제공하고 있으며, 라마, 딥시크, 큐웬 등의 유명 오픈소스 모델이 공개·확산되는 플랫폼이다. 2025년 개발자 수는 500만명 이상으로 약 224만개의 공개 모델과 56만개의 공개 데이터 셋을 기반으로 AI 모델 개발 협업이 가장 활발히 이루어지고 있다[5)].

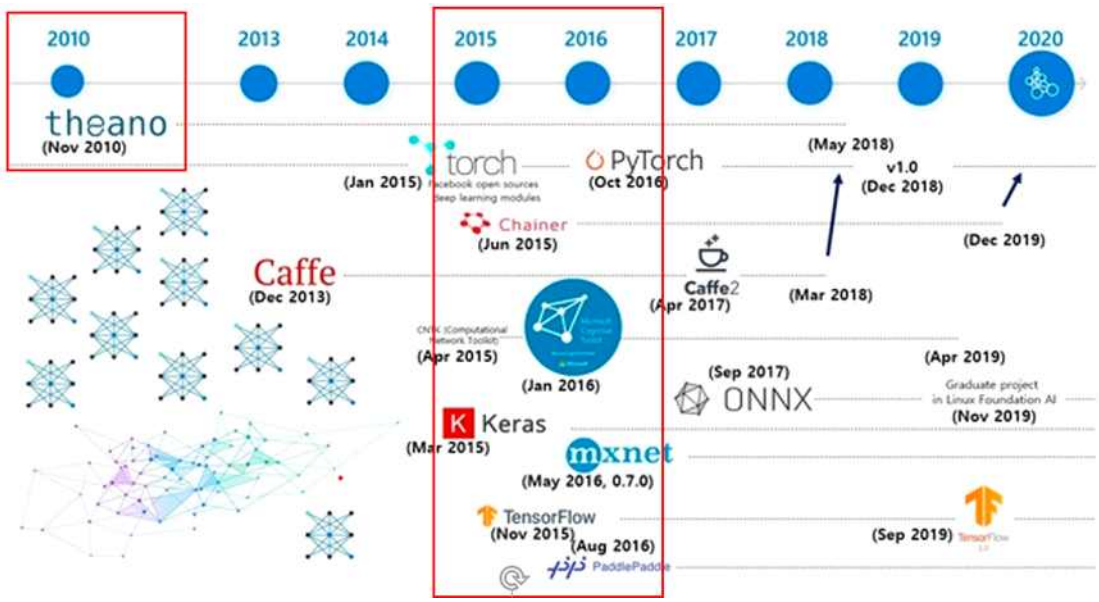
이렇게 오픈소스 생태계가 AI 기술 혁신과 확산을 촉발시킨 주요 계기로는 오픈소스 AI 프레임워크와 허깅페이스(transformer)의 등장이 있다. 2010년대부터 개발되어 공개

3) GitHub, Octoverse 2025, 2025.10.28.
4) 스탠포드 대학, Artifical Intellience Index 2025, 2025.
5) HuggingFace, https://huggingface.co/, 2025.11.21. 방문.

된 오픈소스 AI 프레임워크 기술들이 AI 개발 환경을 표준화하고 개발 편의성를 향상시키며 AI 기술 혁신과 확산의 토대가 되었다. 실제로 오픈소스 AI 프레임워크 기술들은 AI 개발을 위한 표준화된 개발 환경(자동 미분, 텐서 계산, GPU/TPU 연산 등의 기본 기능)과 대규모 학습 환경(분산 연산 및 병렬화, GPU/TPU 클러스터 관리 등)을 API 형태로 제공하여 AI 개발 난이도 완화하며 AI 확산의 원동력이 되었다.

theano(2010년)는 최초의 오픈소스 딥러닝 프레임워크로 딥러닝 환경 구축 편의성을 향상시켰으며, 텐서플로우(2015년)와 파이토치(2016년)는 대규모 학습 기능의 대중화·표준화를 통해 AI 기술 확산과 혁신을 촉발시켰다. 실제로 최근 8개 연구 저널에서 수집한 데이터 분석 결과, 텐서플로우와 파이토치를 사용하는 논문 비중은 약 80% 수준으로 AI 기술 혁신의 근간이 되고 있었다[6]. 또한, EpochAI 보고서에 의하면 AI 학습량을 기반으로 2010년 이전을 딥러닝 이전 시대(2배/20개월 증가), 2010년 이후를 딥러닝 시대 (2배/6개월 증가), 2015년/2016년부터 대규모 시대(딥러닝 시대보다 2~3배 증가)로 구분[7]하고 있는데, 2010년은 최초의 오픈소스 딥러닝 프레임워크 기술인 theano가 공개된 시점이고, 2015년과 2016년은 각각 텐서플로우와 파이토치가 공개된 시점이다.

[그림 5] AI 확산 및 혁신을 촉발한 오픈소스 AI 프레임워크 기술 변화



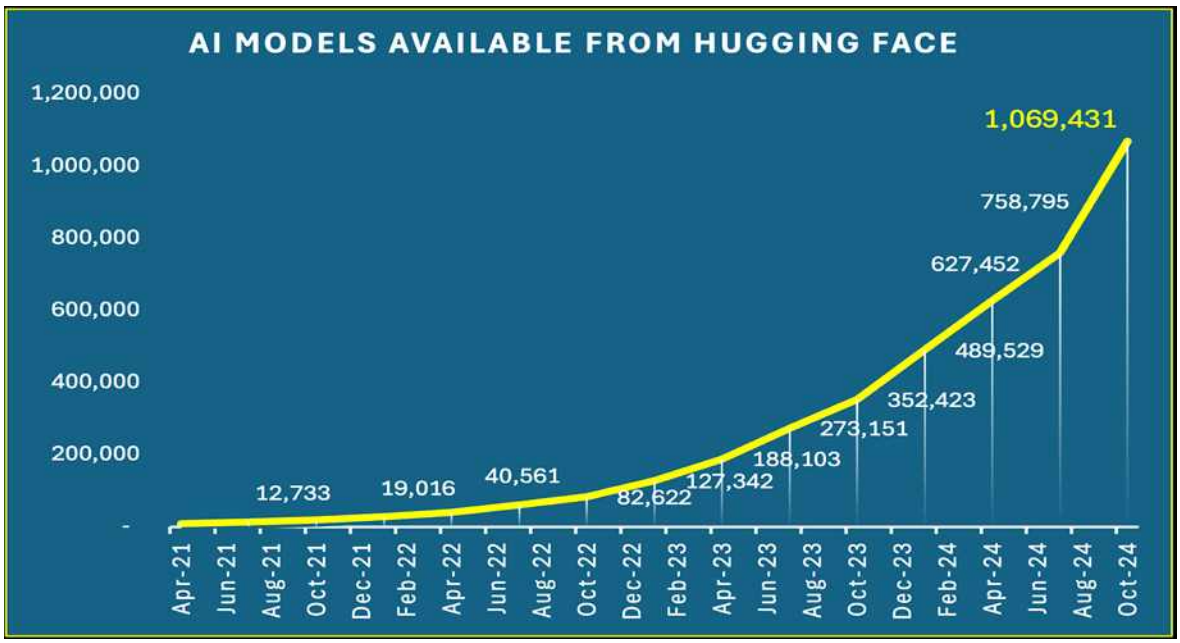
출처) IT 조선, [누구나 개발자 #1] ②오픈소스 AI 개발 도구가 애저클라우드와 만났을 때, 자료 수정

허깅페이스는 머신러닝 모델 개발·배포를 위한 라이브러리(transformers 등)를 오픈

6) Ryan O'Connor, PyTorch vs Tensorflow in 2023, Speech&Text, 2021.12.14.
7) EpochAI, Compute trends across three eras of machine learning, 2022.05.02.

소스로 개발하여 플랫폼 서비스를 제공하고 있다. transformers는 깃허브에서 개발된 오픈소스 프로젝트(스타 수 14만개 이상)로 허깅페이스의 급성장의 핵심 요소로 AI 분야 타 프로젝트(텐서플로우, 파이토치 등) 보다 스타 수가 매우 빠르게 증가하였다[8]. 이렇게 허깅페이스에서 제공되는 AI 개발 환경(transformers, 텐서플로우/파이토치 연계, 엔비디아 칩 가속 기능 등)으로 AI 개발 편의성이 더욱 향상되면서 허깅페이스에서 개발되는 모델 수가 아래 그림처럼 ’21년 만개 이하에서 ‘24년 백만개 이상으로 급증[9]하였고, 최근에는 224만개 이상의 모델들이 공개되어 있다.

[그림 6] AI 확산 및 혁신을 촉발한 오픈소스 AI 프레임워크 기술 변화



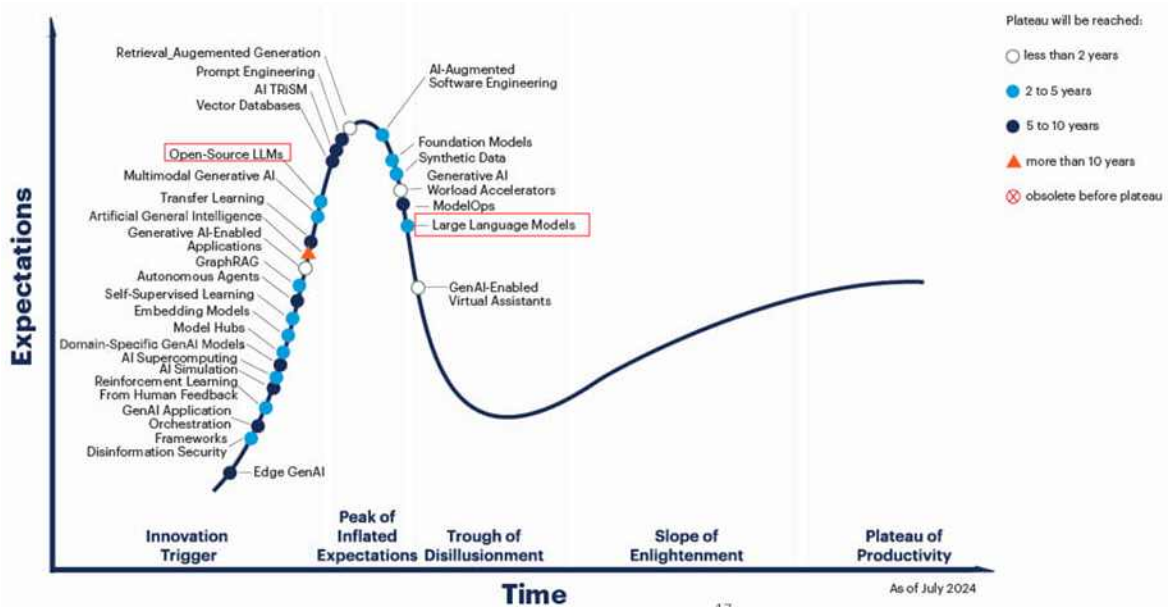
출처) Security Boulevard, Hugging Face Has Become a Malware Magnet, 2024.10.23.

이와 같이 오픈소스 생태계의 AI 분야가 빠르게 성장한 배경에는 AI 산업이 빠르게 성장하며 오픈소스 기반 AI 기술들이 주목받기 때문이다. 글로벌 AI 산업 규모는 글로벌 리서치 기업인 Precedence Research는 2025년 글로벌 AI 산업 규모를 약 7,576억 달러로 예상하고 있으며 연평균 19.2%씩 성장하여 2034년에는 3조 6,804억 달러에 이를 것으로 예상하고 있다[10]. 이는 IMF의 2025년 글로벌 경제 성장률 예측치 3.2%[11] 대비 6배 높은 수치로 글로벌 경제 성장을 AI 산업이 주도함을 알 수 있다.

8) Security Boulevard, Hugging Face Has Become a Malware Magnet, 2024.10.23. https://star-history.com, 2025.11.25.
9) Star History, https://star-history.com, 2025.11.25. 방문
10) Precedence Research, Artificial Intelligence(AI) Market Size, Share and Trends 2025 to 2034. 2025.09.29.
11) 한국무역협회, IMF, 올해 세계성장률 3.2%로 0.2%P↑…“무역합의로 관세영향↓”, 2025.10.15.

이렇게 세계 경제에서 AI 기술 및 산업의 영향력이 커짐에 따라 많은 기업들은 AI 기술 확보를 위해 오픈소스AI에 대한 관심이 증가하고 있다. 2024년 가트너가 발표한 생성형AI 하이퍼사이클에서 대형 언어 모델(LLM)의 기대 심리가 하락하는데 반해 오픈소스 LLM의 기대 심리는 상승하고 있는 것으로 예측하였다.

[그림 7] 가트너의 생성형AI 하이퍼사이클

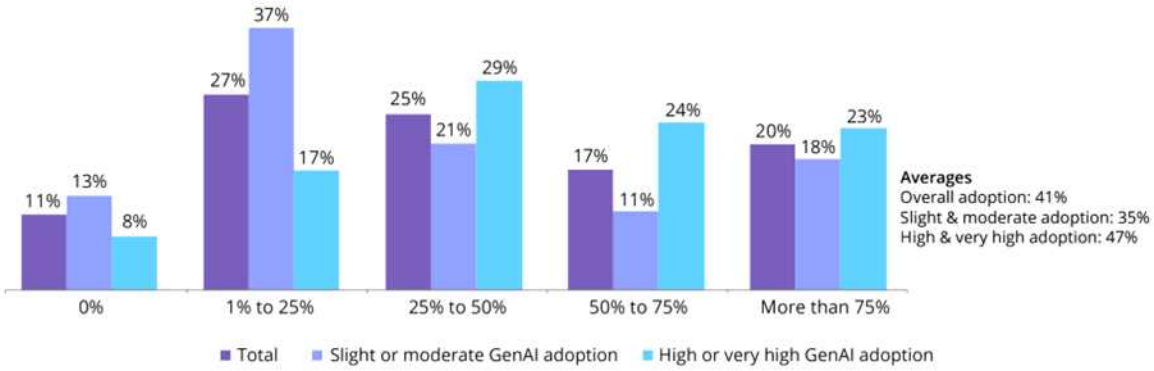


출처) 가트너, 생성형AI 하이퍼사이클 2024. 2024.07.

실제로 많은 기업들이 기술이 공개된 오픈소스 기술을 적극 활용하며 다양한 혜택을 얻기 위해 노력하고 있다. 올해 발표한 리눅스 재단의 보고서인 “The Economic and Workforce Impacts of Open Source AI” 에 의하면, AI 기술을 채택한 글로벌 기업의 89%가 AI 인프라에 오픈소스 기술을 활용[12]하고 있으며, 단지 11% 기업만이 오픈소스를 활용하고 있지 않다고 응답하였다. 그 중에서 자원(인력, 자본 등)에서 상대적으로 부족한 중소기업들이 오픈소스를 보다 적극적으로 활용하고 있었으며, 종사자 규모가 10-249 명 사이인 중소 기업들이 250명 이상인 중균 기업과 만명 이상인 대기업 보다 미활용이 비율이 낮았다[13].

[그림 8] AI 개발 기업 중 오픈소스 채택률

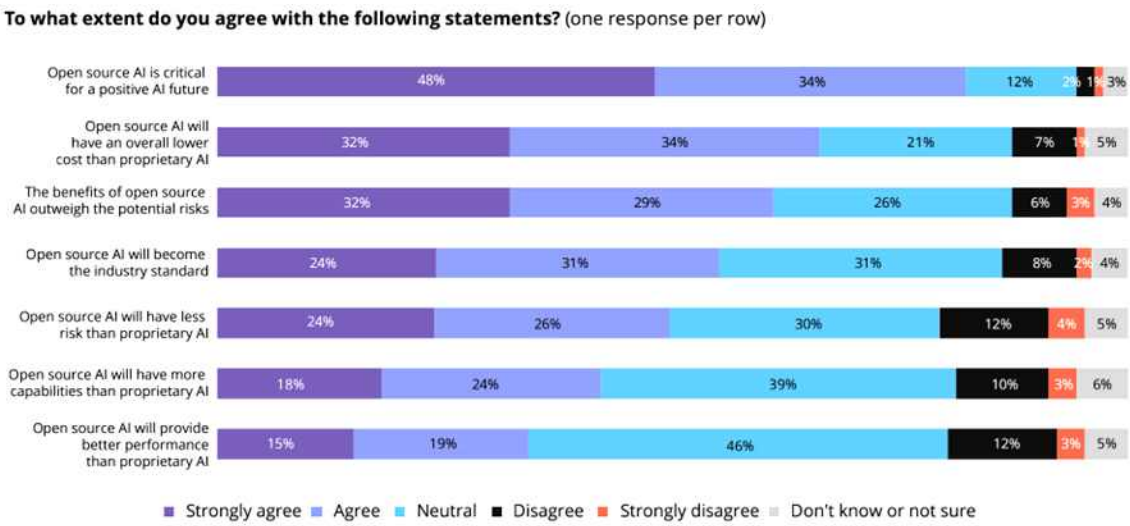
12) 해당 보고서에서 활용하는 광의의 오픈소스AI 개념으로 AI 시스템과 이와 연계된 기술 요소(SW, 데이터, 모델 파라미터, 도구와 문서 등)을 의미함
13) Linux Foundtaion & Meta, The Economic and Workforece Impact of Open Source AI, 2025.05.



출처) 리눅스 재단, The Economic and Workforce Impacts of Open Source AI, 2025.05.

이렇게 AI 개발 과정에 기업들이 오픈소스를 활용하는 이유는 비용 절감, 낮은 위험성, 산업 표준 같은 기업 경쟁력과 직결되는 이슈 때문이며 특히 기업 효율성 측면이 가장 중요한 요인으로 분석되었다. 실제로 응답자의 84%가 오픈소스AI[14]가 미래에 중요해질 것으로 응답하였다. 그리고, 66%의 응답자들이 폐쇄성 AI 보다 낮은 비용으로 제공되며, 61% 응답자들은 폐쇄형 AI 보다 잠재 위험이 적을 것이며, 55%의 응답자들은 산업 표준이 될 것으로 예측하는 응답을 하였다. 이러한 응답 결과는 오픈소스AI의 주요 장점은 낮은 비용, 낮은 잠재적 위험, 산업 표준(호환성) 때문으로 판단된다.

[그림 9] 오픈소스AI의 장점



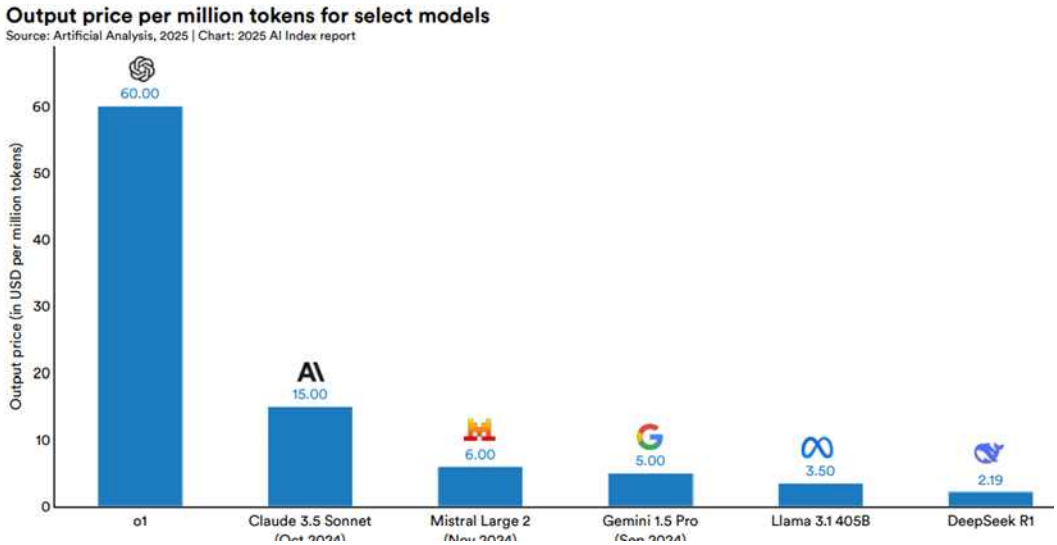
출처) 리눅스 재단, The Economic and Workforce Impacts of Open Source AI, 2025.05.

오픈소스AI(생성 AI 분야의 오픈 모델)가 폐쇄형 AI 모델(상용AI 모델)과 비용 비교는

14) 해당 보고서에서 활용하는 협의의 오픈소스AI 개념을 기반으로 생성 AI 영역의 오픈 모델을 기반의 설문 결과임

다양한 조건들이 있어서 엄격한 비교를 하기 어렵지만, 1차적으로 토큰(데이터 처리 단위)을 기반으로 비교시 최대 1/30(OpenAI와 Deepseek-R1 기준)에 불과하다는 연구 결과가 있다. 그리고, 다른 측면에서 폐쇄형 AI 모델을 활용할 경우에는 클라우드 서비스 형태로 토큰별 비용 지불 혹은 모델 라이선스 비용 지불이 필요하지만 오픈 모델을 활용할 경우에는 자체 구축이 가능해 이러한 비용을 절감할 수 있다.

[그림 10] 상용AI 서비스와 오픈 모델의 비용 비교



출처) 스탠포드 대학 HAI, AI Index 2025. 2025.04.

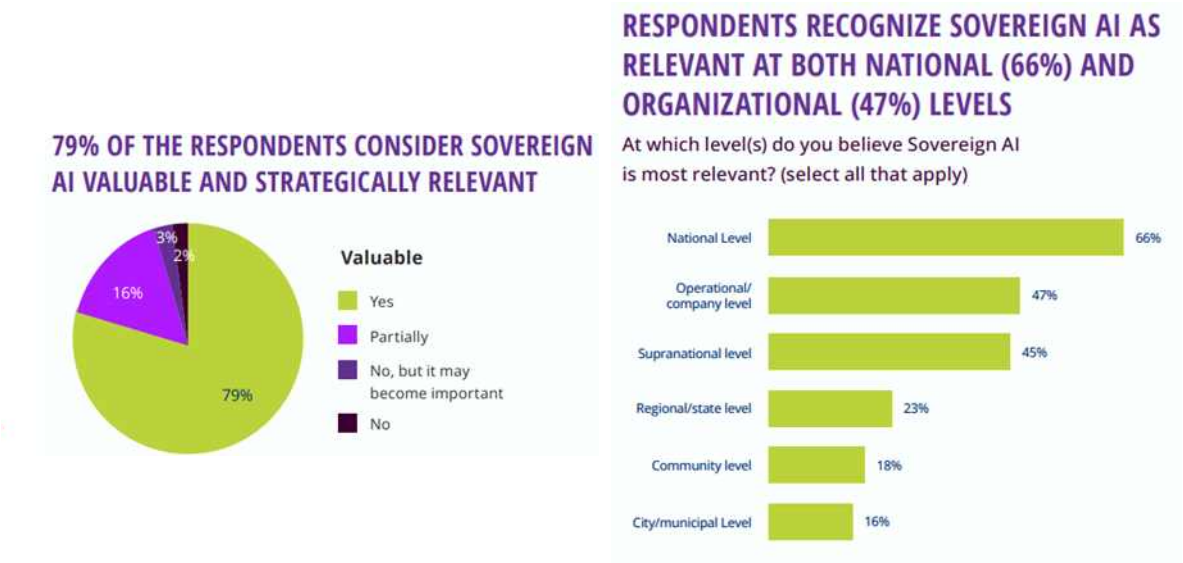
오픈소스AI가 중요해진 다른 이유는 AI 기술·산업 경쟁력 확보를 위한 소버린 AI가 중요해지기 때문이다. 리눅스 재단의 다른 보고서인 "The State of Sovereign AI"에 의하면, 세계적으로 국가 안보 및 경쟁력 강화를 위한 소버린 AI의 필요성에 공감하는 비율이 79%이며, 소버린AI의 대상은 국가(national level, 66%)가 가장 높게 응답되었으며, 이어서 기업과 같은 운영 단위(operational/companay level, 47%), 초국가 수준(supranational level, 45%)로 응답되었다. 이러한 응답 결과는 AI 기술이 진화 중이며, 산업 및 국가 경쟁력에 밀접하기 때문으로 판단된다[15].

실제로 소버린 AI가 필요한 이유에 대한 설문에서 가장 많은 응답은 데이터 유출 방지 및 사생활 보호를 위한 데이터 제어(72%)가 가장 높았으며, 이어서 국가간 분쟁으로 인한 전략자산(AI) 위험성 예방을 위한 국가 보안(69%)과 AI가 국가 산업 경쟁력과 직결됨으로 인한 경제적 경쟁력(48%) 순으로 응답이 많았다. 지역적으로 구분하면 소버린 AI의 필요성은 미국(86%), 유럽(83%), 아시아-태평양(79%) 순으로 응답되었으며,

15) Linux Foundtaion, The State of Soveriegn AI, 2025.08.

이를 위해 82%의 조직에서 자체적인 AI 솔루션을 개발하고 있다고 응답되었다.

[그림 11] 소버린 AI의 중요성



(a) 소버린AI의 중요성 (b) 소버린AI의 대상

출처) 리눅스 재단, The State of Soverign AI, 2025.08.

자체 오픈소스AI 개발 현황은 미국(90%), 유럽(86%)에 비해 아시아-태평양은 (72%)이었으며, 조직 규모 면에서는 종사자 수가 1천명 이상인 대기업(92%)에서 높았으며, 50~1천명(71%), 1~49명(62%) 순으로 낮아졌다. 그리고, 소버린 AI 확보 방안으로 오픈소스 협력이 필요하다는 응답이 94%이었으며 그 이유는 기술의 핵심인 소스코드, 웨이트, 데이터가 공개되어 있어서 모델 웨이트/구조 접근성(84%), 코드 검사・수정(79%), 학습 방법 투명성(76%), 특정 기업 종속성 회피(69%), 파인 튜닝 가능(69%)이 가능하기 때문이었다. 특히 종사자 수가 작은 조직일수록 오픈소스 협업이 중요하다고 인식하였는데, 그 이유는 자원(인적, 자금 등) 부족으로 독자적인 AI 역량(인프라, 데이터, 인력, 모델 등) 확보가 어렵기 때문으로 판단되다.

소버린 AI 확보를 위한 글로벌 오픈소스 협업 대상으로는 가장 가치있는 글로벌 협업 대상은 베이스 및 파운데이션 모델(59%)과 데이터 자원 및 데이터셋(59%)에 이어 도구 및 플랫폼 개발(39%), HW와 연산 인프라(38%) 순서로 응답되었으며, 응답자들이 참여하고 싶은 글로벌 협업의 형태는 오픈소스AI 프로젝트 및 도구 기여(59%), AI 시스템의 기술 표준 제정(45%), 책임있는 AI 원칙 및 사례 협력(45%), 공통 평가 벤치마크 및 지표 개발(40%) 순서로 응답되었다.

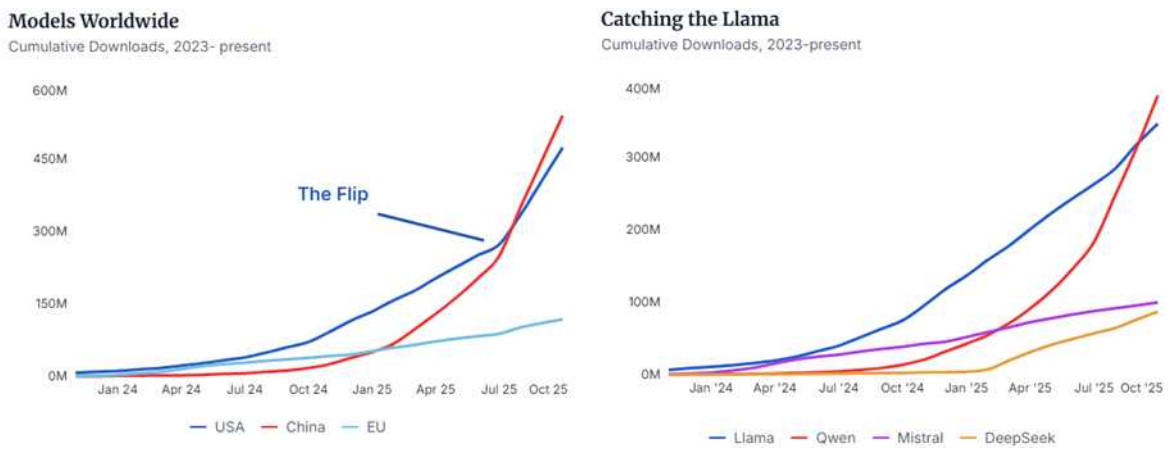
실제로 오픈소스AI 생태계에서 국가적으로 자체 역량 확보 및 글로벌 영향력 확대의

대표적 사례로는 중국 기업들의 오픈소스 모델들이 있으며, 이에 대응하기 위한 미국 민간 단체인 ATOM(America Truely Open Model) 프로젝트[16]가 있다. 전통적으로 오픈소스 생태계는 구글, 마이크로소프트, 인텔(2017년 레드햇 인수), 메타와 같은 미국 기업들이 주도하였고 텐서플로우(구글), 파이토치(메타)와 같은 대표 AI 프레임워크 기술들도 미국 기업들이 공개하면서 오픈소스AI 생태계가 성장해 왔다.

’25년 1월 딥시크가 R1 모델을 공개하면서 중국 AI 기술의 우수성을 공개적으로 입증하면서 오픈소스AI 생태계에서 중국 기업들의 영향력이 빠르게 성장하고 있다. 특히 딥시크 R1은 오픈소스로 공개되면서 준수한 성능(일부 벤치마크에서는 o1보다 좋은 성능을 보임)과 낮은 학습 비용(560만 달러, 약 80억원), 새로운 기술(MoE 구조, 지식 증류 등)으로 전세계적으로 주목받았다. 최근에는 알리바바 큐웬이 빠르게 확산되며, 25년 10월 기준 허깅페이스의 누적 다운로드 수(약 3.9천만 회)에 1위를 기록하며 최근 2년간 누적 다운로드 수 1위이었던 메타의 라마(약 3.5천만 회)를 역전하였다.

그 결과 ‘25년 8월에 중국의 오픈소스 모델의 누적 다운로드 수(허깅페이스)가 미국의 오픈소스 모델의 누적 다운로드 수를 추월하며 중국의 오픈소스 모델이 전세계에서 가장 많이 활용되는 국가가 되었다. 이렇게 중국의 오픈소스 전략은 미국 중심의 오픈소스AI 생태계에 균열을 일으키며 중국의 전세계적인 AI 영향력을 확대하는 중요한 수단이 되고 있다.

[그림 12] 전세계 오픈소스 모델 누적 다운로드 수 현황(허깅페이스 기준)



(a) 주요국 오픈소스 모델의 누적 다운로드 수(허깅페이스) (b) 주요 오픈소스 모델의 누적 다운로드 수(허깅페이스)

출처) ATOM 프로젝트, 2025.08.

이러한 중국의 AI 영향력 확대로 미국에서는 민간을 중심으로 오픈소스AI 생태계 복

16) ATOM Project, https://www.atomproject.ai/, 2025.12.12. 방문.

원 및 오픈소스 모델의 글로벌 주도권 확보를 위해 민간 중심의 투자와 개발 활성화를 위해 ATOM 프로젝트가 ’25년 7월 출범하였다. ATOM 프로젝트는 미국의 대표적인 머신러닝 개발자인 네이선 램버트 주도로 시작된 프로젝트로 미국의 오픈소스 모델 주도권 약화에 따른 중국 오픈소스 모델이 미국내 교육 및 학계에서 영향력을 확대에 따른 우려를 대응하기 위한 목적으로 시작되었다.

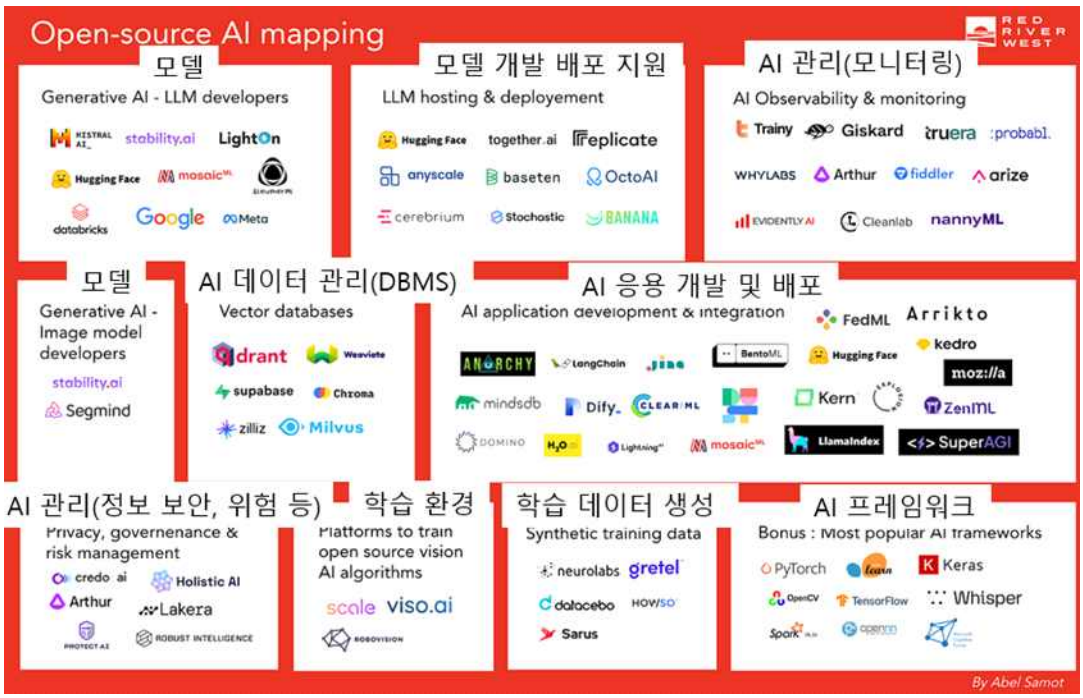
미국 내 일부에서는 미국의 오픈소스 모델 부재로 인한 긴급 보안 이슈(백도어 우려, 생성 코드의 보안 불량, 미국적 가치-자유/평등/독립와 불일치 등)이 증가하는 것을 예방하기 위해 만개 이상의 최첨단 GPU(H100)를 보유한 복수의 오픈소스 모델 연구소 설립하여 개방형 연구 생태계를 복원하고 다양한 규모(100B ~ 600B 이상)의 오픈소스 모델들이 최고 성능으로 제공하는 것을 목표로 하고 있다.

제2절 오픈소스AI 개념의 정립

리눅스 재단 보고서인 “The Economic and Workforce Impacts of Open Source AI”는 오픈소스AI의 개념을 2가지로 구분하고 있다. 첫 번째는 광의의 개념으로 AI 시스템과 이와 연계된 기술 요소(SW, 데이터, 모델 파라미터, 도구와 문서 등)들을 포괄하고 있다. 두 번째는 협의의 개념으로 생성 AI 영역의 오픈 모델만을 의미하는 개념이다. 실제로 오픈소스AI의 개념은 이 2가지 개념들이 혼용되어 사용되고 있다.

첫 번째 오픈소스AI 개념의 범위는 아래 그림과 같이 표현될 수 있다. 이 그림에서는 오픈소스AI는 단순히 생성AI를 위한 LLM만을 의미하지 않고 LLM 호스팅 및 배포, AI 모니터링, 이미지 모델, 벡터 DB, AI 응용 개발 및 통합 환경, 개인정보 및 거버넌스 관리, 학습 플랫폼, 합성 학습 데이터 등과 같이 AI 개발에 필요한 모든 기술적 요소들을 포괄한다. 이 범위를 기반으로 첫 번째 오픈소스AI 개념을 정립한 것이 바로 OSI(Open Source Initiative)의 오픈소스AI 정의(Open Source AI Definition) v1.0이다. 그리고, 두 번째인 협의의 개념인 오픈 모델을 기반으로 개념을 정립화 한 대표 사례로는 리눅스 재단의 모델 개방성 프레임워크(Model Open Framework)가 있다.

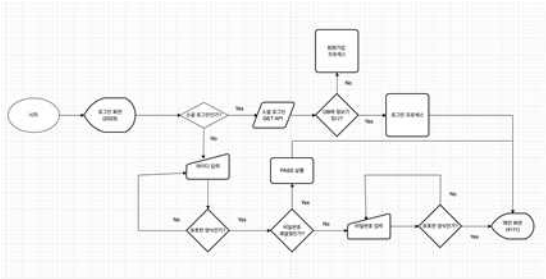
[그림 13] 오픈소스AI 시스템 생태계의 예



출처) Medium, Open-Source AI – Challenges, Opportunities & Ecosystem. 2024.03. 수정

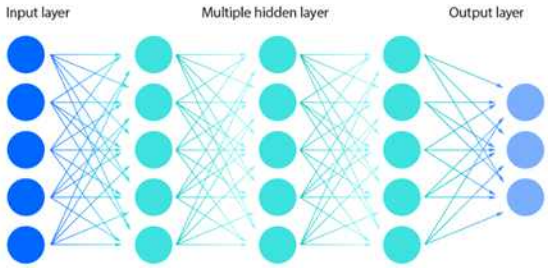
기존 오픈소스 정의(Open Source Definition)[17]가 있음에도 불구하고 이렇게 새로운 오픈소스AI 정의가 필요한 이유는 SW 중심의 오픈소스와 달리 오픈소스AI에는 학습을 위한 데이터, 학습의 결과물인 모델 등이 추가되기 때문이다. 이는 SW와 AI의 논리 구조의 차이 때문에 발생하며 한다. SW는 보편적으로 사람이 현상을 분석하여 도출한 논리(규칙, 절차 등)를 구현한 알고리즘(정의된 논리로 입력 데이터를 반복적으로 처리하는 방식)을 기반으로 하는데 반해 AI는 사람이 구현한 통계적 방법론 알고리즘이 현실 데이터를 학습(분석)하여 통계 및 확률적 규칙과 패턴을 추출하여 이를 인공신경망 형태로 구조화한 논리 구조를 가지기 때문이다.

[그림 14] SW와 AI의 논리 구조 차이



(a) SW 논리 구조의 예 (알고리즘)

출처) UIUX Information



(b) AI 논리 구조의 예(인공신경망 구조)

출처) crone_hi.log

이와 같이 AI는 인경신경망 구조 기반의 논리 구조로 인하여 SW와 달리 데이터, 모델(모델 구조와 웨이트)가 필요하지만, 기존 오픈소스 정의에서는 이에 대한 명확한 공개 기준이 없기 때문에 발생하는 혼동이 발생하게 되었다. 대표적으로 메타의 라마(Llama), 트위터의 Grok, MS의 Phi-2 등은 데이터, 소스코드, 모델을 모두 공개하지 않고 일부만 공개하여 오픈소스의 핵심인 투명성 및 제현성에 제약이 있음에도 오픈소스로 홍보하고 있으며, 이에 대해 일부에서는 오픈소스와 구분되는 오픈워싱(Open washing)이라는 비판이 있다.

따라서, 이러한 문제를 해결하기 위해 오픈소스AI의 구성요소와 공개조건을 명확히 함으로써 오픈소스AI의 자율성, 투명성, 재현성, 재사용, 협력적 개선이 가능하도록 하기 위한 오픈소스AI의 기술적 개념이 필요해졌고, 이러한 기술적 개념을 구체화한 대표 사례로 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 있다.

OSI의 오픈소스AI 정의 v1.0은 약 2년간의 논의를 거쳐 AI 시스템 특성을 고려한 새로운 오픈소스AI 개념을 '24년 10월에 발표하였다. 이 정의는 OECD(Organisation

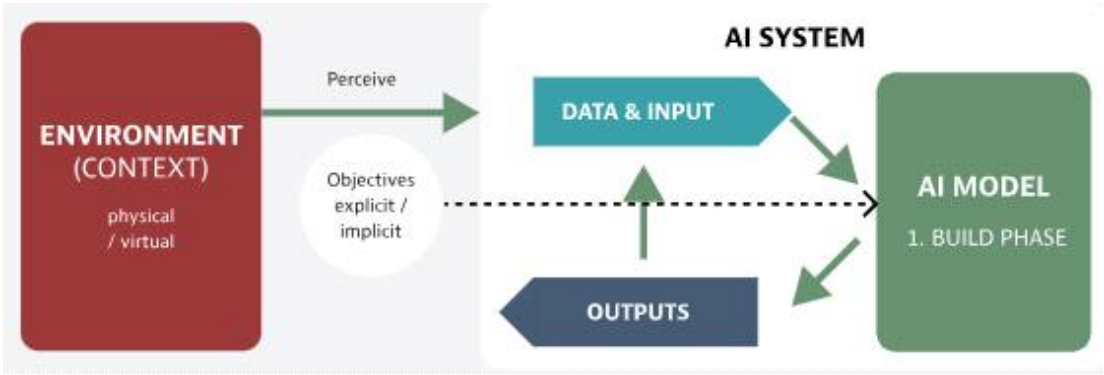
17) Open Souce Initative, Open Source Definition, https://opensource.org/osd, 2025.12.24. 방문.

for Economic Co-operation and Development)의 AI 원칙(Principles)에서 정의된 AI 시스템을 기반으로 정의되어 완전한 AI 기능 구조와 그 구조 내부의 개별 요소 모두를 포괄하는 광의의 오픈소스AI 개념으로 규정되었다.

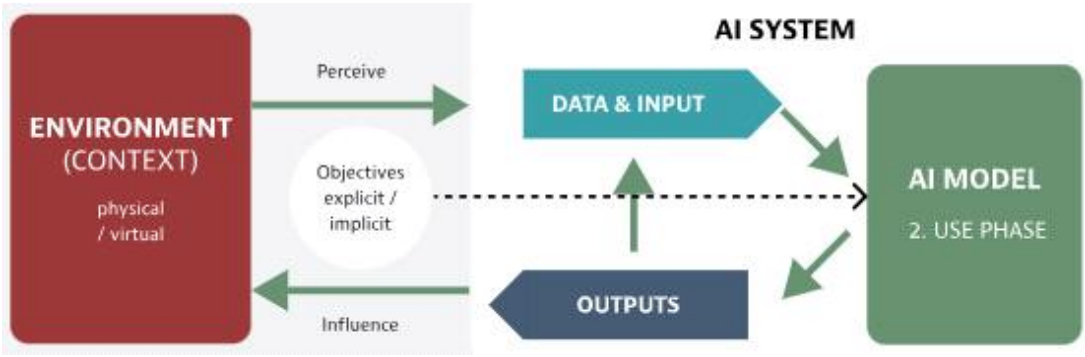
OECD AI 원칙의 AI 시스템은 “명시적 또는 암묵적 목적을 위해 입력된 입력을 바탕으로 예측, 콘텐츠, 추천, 또는 물리적 또는 가상 환경에 영향을 미칠 수 있는 의사결정과 같은 출력을 생성하는 방법을 추론하는 기계 기반 시스템”으로 정의되어 있다[18].

[그림 15] OECD의 AI 시스템 정의

(a) 구축(Build) 단계 AI 시스템 (배포 전)



(b) 사용(Use) 단계 AI 시스템(배포 후)



출처) OECD, OECD AI Principles Overview, https://oecd.ai/en/ai-principles

OSI에서 정의된 오픈소스AI는 자유SW의 4가지 자유(실행, 연구/변경, 복제/배포, 환원)[19]와 유사한 4가지 자유(사용, 연구, 수정, 공유)가 보장된 AI시스템으로 아래와 같이 정의된다. 즉, 오픈소스AI는 목적(상용화, 연구 등)과 대상에 대한 ① 차별과 허가

18) OECD, AI Principles Overview, https://oecd.ai/en/ai-principles, 2025.06.17. 방문
19) GNU 운영체제, 자유 소프트웨어란 무엇인가?, https://www.gnu.org/philosophy/free-sw.ko.html, 2025.11.20. 방문

없이 사용하고, ② 작동 및 구성요소 검사 방법을 연구하며, 출력 변경을 포함한 ③ 시스템을 수정하고, 수정 여부와 관계없이 ④ 공유할 수 있는 AI 시스템을 의미한다.

오픈소스AI 정의 v1.0의 4가지 요구사항(자유)

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.
- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

이러한 오픈소스AI 정의를 구체적으로 이해하기 위한 사례로 머신러닝 시스템을 수정하기 위한 선호되는 구성요소들을 3가지(데이터 정보, 코드, 매개변수)로 구분하며 구체화하고 있다. 비록 오픈소스AI 정의 v1.0를 이해 돕기 위한 다양한 AI 시스템 사례들을 제시하지 못한 점이 아쉽지만 핵심 AI 시스템인 머신러닝 시스템을 사례로 제시한 점은 일부 긍정적이다.

그리고, 아래 표와 같이 데이터 정보, 코드, 매개 변수에 해당하는 공개 항목들을 구체적으로 제시하면서 오픈소스AI 시스템이 되기 위한 공개 항목들을 도출하고 있다. 다만, 데이터와 매개변수는 소스코드에 적용되는 OSI 승인 라이선스를 적용하지 않고, 법적으로 모호한 조건(Term)이라는 용어를 활용함으로써 다소 모호하게 정의하는 한계점을 가지고 있다.

<표 1> 머신러닝 시스템 수정을 위해 선호되는 형태

구성요소	개요	공개되는 정보의 예	제공 조건
데이터 정보	동등한 시스템 구축을 위해 시스템 학습에 사용된 데이터에 대한 충분한 상세 정보	(1) 학습 데이터에 대한 완전한 설명(공유할 수 없는 데이터 포함), 데이터 출처, 범위 및 특성, 수집·선택 방법, 라벨링 절차, 데이터 처리·필터링 방법 (2) 공개적으로 사용 가능한 모든 학습 데이터 목록 및 획득 장소 (3) 제3자로부터 얻을 수 있는 모든 학습 데이터 목록 및 유료를 포함한 획득 장소.	OSI 승인 조건

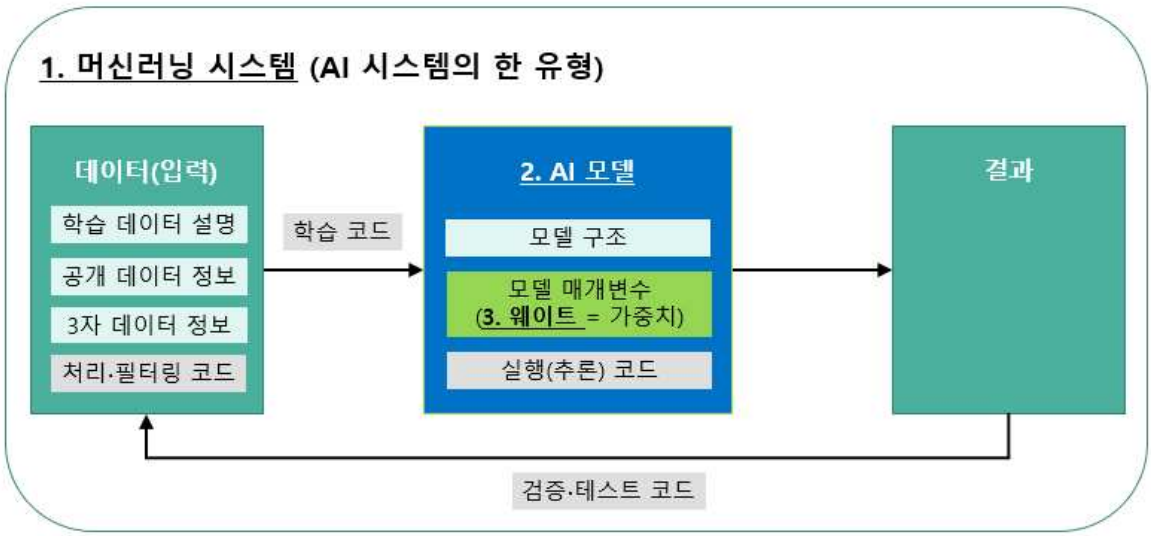
코드	학습 및 실행에 사용되는 완전한 소스코드로 데이터 처리, 필터링 방식, 학습 방식에 대한 전체 명세서	사용된 데이터 처리·필터링 코드, 인수·설정을 포함한 학습 코드, 검증·테스트, 토크나이저 및 하이퍼파라미터 검색 코드와 같은 지원 라이브러리, 추론 코드 및 모델 아키텍처	OSI 승인 라이선스
매개변수	가중치 또는 기타 구성 설정과 같은 모델 매개변수	학습의 주요 중간 단계의 점검 포인트와 최종 최적화 상태	OSI 승인 조건

출처: OSI, Open Source AI Definition v1.0(2024) 내용 SPRi 가공

그리고, 오픈소스AI 정의를 AI 시스템 구성요소에 따른 3가지 기술 범주(① AI 시스템, ② 모델, ③ 웨이트)로 구분하여 3가지 유형의 오픈소스AI 개념을 제시하고 있다. 첫 번째인 오픈소스AI 시스템은 가장 포괄적인 범주로 AI 시스템 전체에 적용된다. 그리고 AI시스템의 하위 구성요소에 따라 독립적 동작이 가능한 AI 모델, 가중치(웨이트)를 별도로 정의하여 복잡한 AI 시스템이 아닌 협의의 오픈소스AI 개념을 위한 AI 모델과 웨이트를 추가로 규정하고 있다.

[그림 16] OSAID의 구성요소에 따른 오픈소스AI 정의의 3가지 범주

(1. AI 시스템, 2 AI 모델, 3. 웨이트)



(자체 작성)

두 번째 개념인 AI 모델은 모델 구조, 모델 매개 변수(가중치 포함), 모델 실행을 위한 추론 코드로 구성된다고 정의하고 있으며, AI 모델의 오픈소스AI 개념인 오픈소스

모델을 규정하였다. 마지막으로 AI 가중치를 주어진 입력으로부터 출력을 생성하기 위한 모델 구조와 연계된 학습된 매개변수 집합으로 정의하면 세번째 오픈소스AI 개념인 오픈 웨이트 개념을 규정하였다.

그리고, 오픈소스 모델과 오픈 웨이트를 추가로 정의하면서 해당 기술 범주인 AI 모델과 AI 가중치 공개와 함께 추가 공개 조건(재현성을 위한 해당 매개변수(가중치)를 도출하는데 사용된 데이터 정보와 코드(학습)들)을 제시하고 있다. 따라서, 오픈소스 모델은 모델 구조, 매개변수(가중치), 실행(추론) 코드와 더불어 매개변수 생성에 활용된 학습 데이터와 학습코드가 같이 공개되어야 한다. 그리고, 오픈 웨이트는 매개변수와 매개변수 실행에 필요한 일부 모델 구조 정보(모델 구조와 연계된 매개변수)와 함께 매개변수 생성에 활용된 학습 데이터와 학습코드가 같이 공개되어야 한다.

하지만, 최근 업데이트된 오픈웨이트(Open Weights) 탭은 현실적으로 통용되는 오픈 웨이트 개념(학습 코드 및 데이터 미공개)과 이에 대한 한계점(재현성 부족, 데이터 불투명성, 규제 장벽, 제한된 협력)에 대해 대해 별도로 제시하고 있다. OSI의 오픈소스 AI 정의는 오픈소스AI의 중요성과 가치에 대한 재인식시켰지만, 개념적 정의로써 다소 모호함과 엄격한 기준에 대한 비판이 일부 제기되고 있다.

비록 24개 단체/기업과 100명 이상의 개인들이 AI의 개방형 혁신을 위한 공식적 논의의 첫걸음으로써 오픈소스AI 정의(v1.0)의 중요성과 가치를 공개 지지[20]하고 있으며, 주요 지지 선언 단체로 모질라 재단, 이클립스 재단, 수세, 오픈인프라 재단, code.gouv,fr(프랑스 오픈소스 공유 허브), OSG-JP(Open Source Group Japan), KAIYUAN(중국 오픈소스 협회) 등과 같은 비영리 단체가 다수를 차지하며 미국 대학(프린스턴, 조지워싱턴, 조지아공대, 카네기 멜론 등)을 공개적으로 지지 선언을 하였고, 우리나라에서는 국가 연구기관(ETRI, AI안전연구소)을 중심으로 개인적 지지 선언이 있었다.

하지만, 오픈소스AI 기술로 알려진 기존 모델 중 다수는 오픈소스AI 정의를 충족하지 못하고 있으며, 일부에서는 이를 오픈소스AI 혹은 오픈소스 모델이 아닌 오픈 워싱이라고 공개적으로 비판하고 있다. 실제로 메타의 라마, 구글의 젬마, X의 Grok, 미스트랄AI의 미스트랄, 딥시크 등은 데이터 정보, 관련 모든 코드 등을 전부 공개하지 않고 일부만 공개하거나 오픈소스AI의 4가지 자유를 모두 허용하지 않고 있다.

그리고, 기존 통용되는 오픈 웨이트(매개변수만 공개) 개념은 제한적 활용(파인튜닝

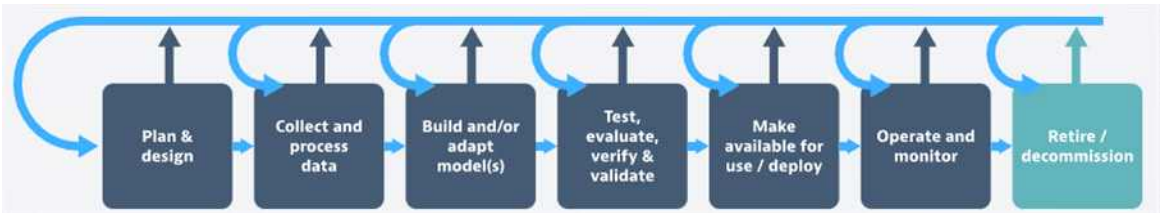
20) OSI, The Open Source AI Definition 1.0 Endorser orginizations, https://opensource.org/ai/endorsements, 2025.06.17. 방문

가능)과 한계점(재현성 부족, 데이터 불투명성, 규제 장벽, 제한된 협력) 때문에 오픈소스AI로 분류될 수 없지만, 부분적 자유가 허용되기 때문에 오픈소스AI 정의에서 가장 논쟁의 요소가 많이 있다. 현재 OSI는 새로운 오픈웨이트 탭을 추가하여 이러한 이슈들을 소개하고 있으며 향후 OSI가 어떻게 대응할지 귀추가 주목되고 있다.

OSI의 오픈소스AI 정의가 AI 시스템 관점에서 큰 기술적 범주에서 규정되어 있지만, 2024년 12월에 발표된 리눅스 재단의 모델 개방성 프레임워크는 오픈소스 모델의 개방성 수준을 분류 방법을 세부적으로 규정하고 있다. 따라서, 모델 개방성 프레임워크는 개념 중심의 OSI의 오픈소스AI 정의와 달리 모델 개발 수명 주기 관점에서 모델 구성 요소들과 적용된 라이선스를 기반으로 개방성을 실질적으로 평가하기 위한 프레임워크를 제공한다.

AI 모델은 SW 개발 방식과 달리 반복적인 데이터 학습을 통해 기계(컴퓨터)가 모델 내부의 최적 웨이트(통계/확률 값) 도출하는 방식으로 개발된다. OECD는 AI 모델의 상위 개념은 AI 시스템 생명 주기를 아래 그림과 같이 정의하고 있다[21]. 우선 계획 수립과 설계를 통해 개발 범위를 결정하고, 그 다음은 모델(AI 시스템) 학습을 위한 데이터 수집 및 처리 과정을 거쳐 모델 논리 구조의 핵심인 인공신경망 구축을 위한 모델 학습을 진행하게 된다. 그리고, 평가, 검증을 통해 적정 수준의 성능과 동작 결과를 검증 후에 배포 과정을 거쳐 운영과 모니터링 단계로 넘어가게 된다. 그리고, 더 이상 사용이 필요없게 되면 폐기 단계로 시스템에서 분리된다.

[그림 17] OECD의 AI 시스템 생명 주기(Lifecycle)



출처: OECD, AI Ssytem lifecycle

모델 개방성 프레임워크는 이러한 AI 모델의 생명 주기 관점에서 모델 재현성과 투명성을 보장하기 위한 구성 요소로 3가지 형태(코드, 데이터, 문서)로 구분되는 총 17개의 구성요소를 도출하였으며, 3가지 형태에 따라 적용되는 개방형 라이선스들을 구

21) OECD AI Principles overview, AI Ssytem lifecycle, 2025.11.20. 방문

분하여 규정하고 있다. (소스)코드 형태로 공개되는 6개 구성요소들은 오픈소스 라이선스를 적용하고, 데이터 형태로 공개되는 6개 구성요소들은 오픈데이터 라이선스를 적용하며, 문서 형태로 공개되는 5개 구성요소들은 오픈 컨텐츠 라이선스를 적용하고 있다.

[그림 18] 모델 개방성 프레임워크의 17개 구성 요소와 적용 라이선스 구분

출처):Linux Foundation 자료(2024) SPRi 재구성

그리고, 17개 구성 요소들의 공개 여부에 따라 3단계(클래스 1: 오픈 사이언스 모델, 클래스 2: 오픈 툴링 모델, 클래스 3: 오픈 모델)로 구분되어 모델 개방성을 평가하고 있다. 개별 구성요소의 개방성은 적용된 개방형 라이선스(OSI 오픈소스AI 정의 v1.0의 4가지 자유 - 사용, 연구, 수정, 공유가 적용된)를 기준으로 판단하고 있다. 추가적으로 3단계 모델 개방성을을 기반으로 오픈소스 모델 사용자를 위한 수준별 모델 활용 가능 범위를 구체적으로 제시하고 있다는 점에서 매우 의미 있다.

<표 2> 모델 개방성 프레임워크의 3단계 개방성 수준

MOF 개방 수준	구성요소 개방 조건	유형	활용 범위
클래스 3. 오픈 모델 (Open Model)	1. 모델 구조 2. 최종 모델 매개 변수 3. 기술 보고서 또는 연구 논문 4. 평가 결과 5. 모델 카드	코드 데이터 문서 문서 문서	• 제약 없는 사용 범위(접속, 사용, 변경, 재배포) • 제품 및 서비스 개발

	6. 데이터 카드 7. 모델 결과 샘플(선택)	문서 데이터/코드	• 미세 조정 및 정렬 • 모델 최적화
클래스 2. 오픈 도구 모델 (Open Tooling Model)	• 클래스3의 모든 구성요소 8. 훈련, 검증 및 테스팅 코드 9. 추론 코드 10. 평가 코드 11. 평가 데이터 12. 지원 라이브러리 및 도구	코드 코드 코드 데이터 코드	• 학습 과정의 이해 • 벤치마크 결과 검증 • 추론 최적화
클래스 1. 오픈 사이언스 모델 (Open Science Model)	• 클래스2와 3의 모든 구성요소 13. 연구 논문 14. 데이터셋 15. 데이터 전처리 코드 16. 중간 단계 모델 매개변수 17. 모델 메타데이터 (선택)	문서 데이터 코드 데이터 데이터	• 모든 단계의 분석과 검사 • 유사 모델의 재구축 • 데이터 탐색 및 실험

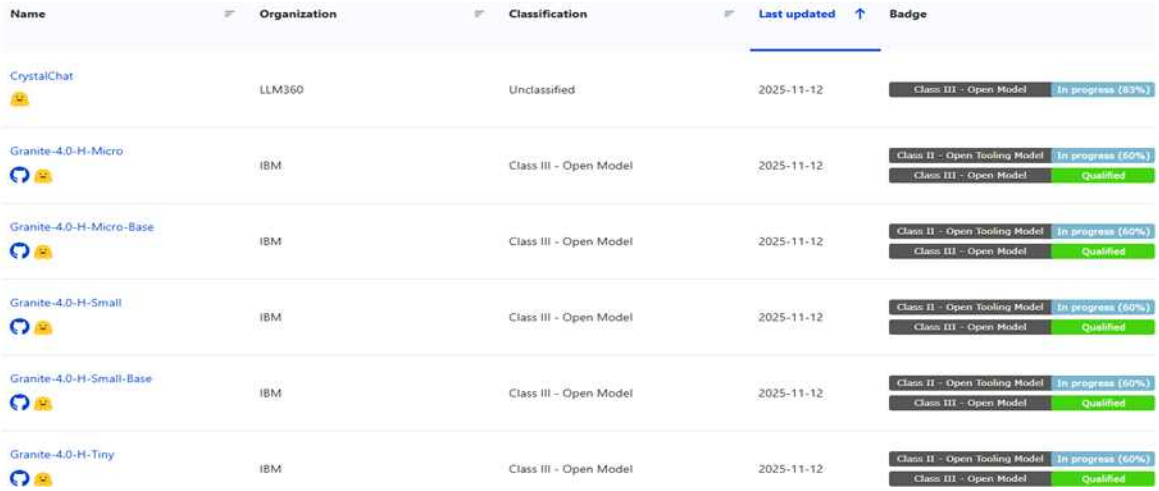
출처: 리눅스 재단, 모델 개방성 프레임워크 편집

또한, 선언적인 OSI의 오픈소스AI 정의와 달리 주요 모델들의 개방성을 실질적으로 평가하기 위한 모델 개방성 도구(Model Openness Tool)과 오픈소스 모델의 구성 요소인 데이터와 문서에 적용 가능한 개방형 라이선스 정보들을 같이 제공한다는 점에서 보다 실용적이다. 모델 개방성 도구는 모델 개방성 프레임워크의 실질적 활용을 위해 구성요소별 공개 수준을 정량화하여 모델 개방성을 실질적으로 평가*하는 도구이며, 현재 베타 버전이 제공되고 있다[22]. 이 도구를 통해 모델 개방성 프레임워크가 단순 개방성 유형(클래스) 구분이 아니라 개방성 수준을 정량적으로 평가함으로써 모델간 개방성 수준 비교가 가능하게 한다.

또한, 오픈소스 모델에 적용가능한 개방형 라이선스 약 700개를 수집하여 OSI 승인, 자유소프트웨어 재단 인증 여부 정보를 별도로 제공함으로써 모델 사용자에게 위험성을 직접적으로 판단할수 있게 해준다[23]. 이는 OSI의 오픈소스AI 정의에서 학습 데이터와 모델 구조에서 OSI 승인 조건이라는 애매모호한 표현을 사용한 것과 달리 더 구체적이다.

22) 리눅스 재단, Model Openness Framework, Models, https://mot.isitopen.ai/models, 2025.11.20.
23) 리눅스 재단, Model Openness Framework, Liceses, https://mot.isitopen.ai/licenses, 2025.11.20.

[그림 19] 모델 개방성 도구의 평가 결과의 예



Name	Organization	Classification	Last updated	Badge
CrystalChat	LLM360	Unclassified	2025-11-12	Class III - Open Model: In progress (83%)
Granite-4.0-H-Micro	IBM	Class III - Open Model	2025-11-12	Class II - Open Tooling Model: In progress (60%); Class III - Open Model: Qualified
Granite-4.0-H-Micro-Base	IBM	Class III - Open Model	2025-11-12	Class II - Open Tooling Model: In progress (60%); Class III - Open Model: Qualified
Granite-4.0-H-Small	IBM	Class III - Open Model	2025-11-12	Class II - Open Tooling Model: In progress (60%); Class III - Open Model: Qualified
Granite-4.0-H-Small-Base	IBM	Class III - Open Model	2025-11-12	Class II - Open Tooling Model: In progress (60%); Class III - Open Model: Qualified
Granite-4.0-H-Tiny	IBM	Class III - Open Model	2025-11-12	Class II - Open Tooling Model: In progress (60%); Class III - Open Model: Qualified

출처) Linux 재단, Model Openness Framework

[그림 20] 모델 개방성 도구의 개방형 라이선스 정보 형태

Name	License ID	Content Type	OSI Approved	FSF Lib
3D Slicer License v1.0	3D-Slicer-1.0		No	No
3dfx Glide License	Glide		No	No
Abstyles License	Abstyles		No	No
Academic Free License v1.1	AFL-1.1	code	Yes	Yes
Academic Free License v1.2	AFL-1.2	code	Yes	Yes
Academic Free License v2.0	AFL-2.0	code	Yes	Yes
Academic Free License v2.1	AFL-2.1	code	Yes	Yes
Academic Free License v3.0	AFL-3.0	code	Yes	Yes
Academy of Motion Picture Arts and Sciences BSD	AMPAS		No	No

출처) Linux 재단, Model Openness Framework

이와 같이 모델 개방형 프레임워크는 OSI의 오픈소스AI 정의와 유사하게 자유SW의 4가지 자유를 기반으로 오픈소스AI 개념을 정의하고 있지만, 오픈소스AI 정의가 AI 시스템 관점에서 기술적 범주를 가지고 3가지로 구분한 것과 달리 AI 모델의 개방성 정도를 기준으로 3가지 유형으로 구분한 점에서는 차이가 있다. 또한 개방성 수준을 구분하기 위해 17개의 구성요소를 상세히 규정하고 있으며, 적용 라이선스도 명확하게 정의하고 있다.

제3절 요약 및 시사점

1. 요약

오픈소스 생태계는 개발자 중심의 SW 개발 협업을 위해 출발하여 상용SW 대안(운영체제, DB, 웹서버 등)으로 주목을 받으면서 플랫폼 SW(모바일, 인공지능, 빅데이터, 클라우드 등) 혁신을 추도하며 성장하였다. 이 과정에 오픈소스 AI 프레임워크 기술들이 머신러닝 개발 편의성과 효율성을 제공하면서 본격적인 머신러닝 중심의 인공지능 시대를 개막하게 된 동력이 되었으며, 텐서플로우(2015)와 파이토치(2016)는 머신러닝의 대형화를 선도하며 본격적인 대형 언어 모델(LLM) 시대의 개막을 촉진하며 신 AI 혁신과 시장 창출의 촉매제 역할을 하였다.

2020년대 들어서 오픈소스 생태계는 AI 기술 분야를 중심으로 지속적으로 성장하고 있다. 대표적으로 핵심 오픈소스 개발 협업 플랫폼인 깃허브는 2024년에 개발자 수가 3600만명이 증가하였으며, 4320만 건의 기여(풀리퀘스트)가 발생할 정도로 매우 활성화되어 있다. 그 중심에는 약 430만개의 AI 관련 프로젝트들이 있으며 이들의 누적 스타 수도 1764만개로 매우 빠르게 증가하고 있다. 또한 핵심 AI 개발 협업 플랫폼인 허깅페이슨도 2025년 가입자 수가 500만명을 넘어섰고, 공개된 모델 수가 224만개가 넘어서며 빠르게 영향력을 키우고 있다.

이렇게 오픈소스AI(오픈소스 생태계의 AI 기술 분야)가 빠르게 성장하는 배경에는 많은 개발자들이 AI 개발에 오픈소스를 보편적으로 활용하고 있기 때문이다. 리눅스 재단 보고서에 의하면 AI 개발 기업 중 인프라에 오픈소스 채택 비율이 89%일 정도로 오픈소스는 AI 기술 인프라 역할을 제공하고 있으며, 그 이유는 비용 절감, 낮은 위험성, 산업 표준 같은 이유들 때문이었다. 실제로 일부 연구에서 오픈 모델이 상용AI 서비스와 비교시 토큰 비용이 1/30에 불과하다는 연구 결과가 있다.

또한, 국가적 차원의 AI 역량 강화를 위한 소버린 AI 관점에서 오픈소스AI가 중요하다는 의견이 있다. 실제로 리눅스 재단 보고서에서 국가 안보 및 경쟁력 강화를 위한 소버린 AI의 필요성에 공감하는 비율이 79%이었으며, 국가적 측면에서 중요성(66%)이 가장 높게 응답되었다. 그 이유는 AI 기술 혁신이 가속화되면서 국가 경제 및 산업 활성화에 큰 영향을 미치고 있으며, 학습용 데이터 유출에 따른 국가 안보 및 기업 경쟁력 훼손이 우려되기 때문에 오픈소스AI를 활용하여 자체 AI 시스템 구축 역량 확보가 중요해지고 있다.

이렇게 오픈소스 생태계에서 AI 기술 분야의 비중이 증가하고 영향력이 확대되면서 이를 오픈소스AI라고 부르며 기존 오픈소스 생태계와 구분되는 용어가 빈번하게 사용되고 있다. 또한 기존 SW 중심의 오픈소스 생태계와 구분하여 AI의 기술적 특징을 기반으로 오픈소스AI의 재현성과 투명성을 제고하기 위한 오픈소스AI 개념이 새로이 정립되고 있다. 대표적으로 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 있다.

OSI의 오픈소스AI 정의는 4가지 자유(사용, 연구, 수정, 공유)가 보장된 AI 시스템으로 정의하고 있으며, 기술 범주에 따라 AI 시스템, AI 모델, AI 웨이트로 구분하며 각각에 대한 오픈소스AI 시스템, 오픈소스 모델, 오픈 웨이트로 구분하고 있다. 이에 반해 리눅스 재단의 모델 개방성 프레임워크는 AI 개발 생애주기 관점에서 모델 재현성을 제공하기 위한 개방성 수준을 정의하고 있다. 가장 개방적인 오픈사이언스 모델, 오픈 도구 모델, 오픈 모델로 구분하여 구분된 모델 개념간 활용 범위를 명확히 제시하고 있다.

2. 시사점

최근 부상하고 있는 오픈소스AI 생태계와 새로운 오픈소스AI 개념을 중심으로 다음과 같은 시사점들을 제시한다.

1) 오픈소스AI의 중요성 : AI 기술·산업 혁신의 원동력

최근 AI 시장의 급성장 배경에는 생성형 AI 기술이 있다. 과거와 달리 OpenAI가 chatGPT-3를 공개하며 많은 사람들은 GPT-3 기술의 발전에 놀라워하며 문서 요약, 다국어 번역, SW 코드 생성, 이미지 생성 등 다양한 분야에 적용하며 생산성 혁신 가속화하고 있다. 이러한 chatGPT-3 기술·산업 혁신의 근간에 많은 오픈소스 기술들이 있다.

2010년 공개된 오픈소스 AI 프레임워크 기술인 theano는 머신 러닝 시대를 본격화시킨 대표적인 오픈소스 기술로 현대적 머신러닝 프레임워크를 기술적 토대를 제공하고 있다. theano의 자동 미분, GPU 활용, 연산 병합 등의 기술들은 CNN, RNN 같은 인공신경망 구조 활용을 확산시켰다. 그리고 2015년 공개된 텐서플로우와 2016년 공개된 파이토치는 theano 기술을 확장 시켜 텐서라는 단위 연산 기능을 구축하고, GPU 최적

화, 파이썬 활용, 디버깅 강화 등을 통해 대규모 머신러닝 시대를 개막시켰다.

이후 머신러닝 시대가 본격화되고 대형화가 확산되면서 다양한 머신러닝 구조에 대한 오픈소스 기술들(트랜스포머, BERT, GPT-1/2 등)이 공개되면서 이들의 핵심 기술들이 GPT-3에 반영되며 생성AI 시대를 개막하는 근간이 되었으며, AI 기술 혁신을 넘어 신산업 태동 및 기존 산업 혁신을 촉발하는 근간이 되었다.

2) 다양한 관점의 오픈소스AI의 전략적 가치 : ① 기술 인프라, ② 기업 경쟁력 요소, ③ 소버린 AI 수단, ④ 기술 경쟁 수단, ⑤ 글로벌 협업 수단

이렇게 오픈소스 기술들이 AI 기술·산업 혁신을 촉발하는 근간이 되면서 AI 기술 개발에 오픈소스 기술들이 널리 활용되고 있다. DB 분야의 오픈소스 그래프 DB, 개발 환경에 텐서플로우, 파이토치. 트랜스포머스(허깅페이스), 학습 데이터 생성 등 AI 개발에 필요한 다양한 기술 요소들이 오픈소스로 개발되고 있다. 이들은 광의의 오픈소스 AI 개념으로 묶이면서 AI 개발을 위한 기술적 인프라 역할을 담당하고 있다.

오픈소스 기술들이 AI 개발 과정에 보편적으로 활용되면서 기업의 AI 개발 생산성에 오픈소스 기술들이 직접적인 영향을 주고 있다. AI 개발에 필요한 모든 기술들을 자체 개발할 필요없이 우수한 오픈소스 기술들을 도입하게 되면 비용 절감, 개발 기간 단축, 기술 위험 회피, 산업 표준(호환성) 등에서 장점이 있기 때문이다. 따라서 광의의 오픈소스AI 기술들은 기업 경쟁력의 핵심 요소로 부각되고 있다. 특히 인적·재정적 자원이 풍부하지 못한 중소기업들은 대기업 보다 오픈소스AI에 대한 더 많은 관심을 보이고 있다.

또한, 오픈소스AI는 기업 뿐만 아니라 국가적 차원의 안보 및 경쟁력 강화의 수단으로 인식되고 있다. 실제로 오픈소스AI가 AI 경쟁력 확보 수단 뿐만 아니라 데이터 제어, 국가간 분쟁에서의 AI 주권 확보, 국가 보안, 산업 경쟁력 측면에서 중요한 영향을 미치기 때문이다. 따라서, AI 기술 역량 확보 및 자체적인 AI 시스템 확보를 위해 오픈소스AI 기술 내재화가 중요해지고 있다.

최근 미국에서는 최근 부상하는 중국 오픈소스 모델에 대한 대응으로써 미국내 오픈소스 개발 협업을 촉구되는 움직임이 있을 정도로 오픈소스AI는 기술 경쟁 수단화되고 있다. 전통적으로 오픈소스AI 생태계는 미국 기업(구글, 메타, MS 등)들이 새로운 기술을 공개하며 사실상 시장 표준 역할을 담당하며 기술 혁신을 선도해 왔다. 하지만, 2025년 들어서 딥시크의 우수한 성능과 효율성을 기반으로 전세계적으로 주목받고

알리바바의 큐웬이 우수한 성능과 자유로운 활용을 허용해 주면서 중국 모델의 영향력이 급격히 확산되고 있다. 이에 미국은 민간 협업을 활성화하여 지속적인 기술 주도권 확보를 위한 ATOM 프로젝트 시작되면서 오픈소스AI가 기술 주도권 확보 수단으로 여겨지고 있다.

오픈소스 생태계는 SW개발의 개방형 협업을 위해 탄생되었으며 오픈소스의 발전으로 투명한 공개 기술을 활용한 비즈니스 협업 수단으로 고도화되었다. 오픈소스AI 기술들도 투명하게 공개되어 자유로운 기술 협업이 가능하다. 실제로 텐서플로우/파이토치 등은 공개된 개발 협업으로 지속적인 기술 혁신을 통해 사실상 표준 머신러닝 개발 플랫폼이 되었으며, OpenAI의 GPT도 기존 트랜스포머, BERT 등의 기술 요소들이 적용하였고, GPT-1/2의 주요 기술들도 메타 라마 등의 다른 모델에도 적용되어 있다. 이와 같이 오픈소스 기술들은 기술 개발 협업에 매우 유리하기 때문에 글로벌 협업 수단으로 각광받고 있다.

3) 소버린AI를 위한 오픈소스AI의 역할 : ① AI 요소 기술, ② 외부 의존성 완화(자율성 확보), ③ 신뢰성 확보 수단

AI 기술 역량 확보와 독자적인 AI 시스템 운용을 위해서는 오픈소스 기술 활용이 필수 불가결하다. 그 이유는 AI 개발을 위해서는 모델 뿐만 아니라 기반이 되는 개발 환경, 데이터 수집 관리, 성능 검증, AI 디버깅, 배포, 유지 보수 등 다양한 기술적 요소들이 필요하기 때문이다. 그러나 이러한 기술적 요소들 모두를 자체 개발하는 주체는 전세계 어디에도 없기 때문에 1차적으로 AI 관련 다양한 오픈소스 기술들이 반드시 필요하다.

이에 반해 상용AI 서비스를 활용할 경우에는 기술 제공사가 제공하는 기능을 일정 비용을 지불하고 사용할 수 있다. 그리고 사용자 원하는 기능을 제공하지 않는다면 이에 대한 개선 요구를 할 수 있지만 선택권은 기술 제공사에게 있게 된다. 또한 기술 제공사의 비용 인상 또는 서비스 중단이 발생할 경우에는 이러한 외부 변화를 일방적으로 수용할 수 밖에 없다. 따라서 국가 혹은 기업의 전략 분야의 외부 의존성을 완화하기 위해서는 독립적인 AI 역량 확보가 필요하며 이를 위한 가장 우선적인 대안이 바로 기술이 공개된 오픈소스 기술들이다.

그리고, 오픈소스 기술의 특징은 기술의 핵심인 소스코드, 웨이트 등이 공개되어 있어 투명한 기술 검증이 가능하다는 장점이 있다. 상용AI 서비스의 경우 입력(데이터)를

제공하면 그에 따른 결과만 제공하기 때문에 내부적으로 어떠한 판단 근거가 이루어졌는지 알 수 없다. 하지만, 오픈소스 기술들을 활용하면 소스코드에 악성 코드가 있는지 투명하게 검사할 수 있으며, 모델 검증을 통해 사전에 철저하게 검증할 수 있다.

4) 정립중인 오픈소스AI 개념

오픈소스AI 기술들이 전세계적으로 주목받으며, 관련 생태계가 지속적으로 성장하고 있지만, 오픈소스AI 개념은 작년까지 명확하지 않았다. 하지만 작년 말에 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 발표되며 오픈소스AI의 개념이 구체화되고 있다. 특히 자유SW의 4가지 자유를 기반으로 사용, 연구, 수정, 공유라는 오픈소스AI의 핵심 특징이 정립되었다.

하지만, 이들 개념 사이에서 차이가 있다. 우선 OSI의 오픈소스AI 정의는 AI 시스템 관점에서 정립되었으며, 오픈소스AI 시스템, 오픈소스 모델, 오픈 웨이트 개념을 구체화하였다. 그리고 리눅스 재단의 모델 개방성 프레임워크는 AI 개발 생애 주기 관점의 오픈소스 모델의 개방성 수준을 오픈 사이언스 모델, 오픈 도구 모델, 오픈 모델로 구분하여 정립되었다. 그럼에도 이들은 산업계에서 통용되는 오픈 웨이트 모델에 대해서는 구체화하지 않아 보편적 개념과 다소 차이를 보이고 있다. 물론 과거에도 오픈소스 정의가 보편화되는데 걸린 시간을 고려하면 아직 개념이 정립된 초창기라서 이런 혼동을 일시적일 수 있으며, 향후에 현재 정립된 오픈소스AI 개념이 어떻게 변화할지 지켜볼 필요가 있다.

제3장 국내외 오픈소스AI 동향

제1절 현황 조사 개요

최근 글로벌 AI 기업들은 오픈소스 기술 및 생태계를 비즈니스 전략의 수단으로 활용하고 있다. 대표적으로 메타는 OpenAI의 chatGPT의 시장 집중을 견제하기 위해 라마를 전략적으로 공개하고 이후 활용 대상을 점진적으로 확대하였다. 그리고 2025년에 딥시크는 R1 모델을 공개하며 준수한 성능과 비용 우월성을 입증하며 시장에서 영향력을 확보했으며 알리바바는 오픈소스 모델인 큐웬을 통해 시장 장악력을 확대하고 있다.

이러한 글로벌 AI 기업들의 오픈소스 전략을 보면 기존 SW 생태계에서 지난 30년간 다양한 일들이 빠르게 진행되고 있는 것으로 보인다. 초기 오픈소스 기술들은 자유로운 활용을 보장으로 인한 무료 사용이라는 인식이 강했지만, 기업들이 오픈소스 생태계를 장악해 나가면서 오픈소스는 더 이상 무료 공개가 아닌 생태계 확장(고객 확보)과 수익 창출을 위한 전략적 비즈니스 수단으로 자리잡았다.

현재 메타의 라마 시리즈(버젼 1 ~ 4 까지)는 10억 이상의 다운로드를 기록하며 글로벌 생태계에 큰 영향을 끼치고 있다. 2025년 공개된 딥시크의 낮은 학습 비용(600만 달러)은 NVIDIA 시가총액을 846조원이 감소할 정도로 AI 생태계에 큰 영향을 주고 있다. 그리고 한동안 모델을 공개하지 않던 OpenAI도 자체 모델 기술을 gpt-oss-120b를 공개하면서 생태계 영향력을 유지하기 위해 노력하고 있다. 또한 2025년 10월 알리바바의 큐웬은 허깅페이스 누적 다운로드 수에서 메타 라마를 추월하며 전세계적인 주목을 받고 있다. 그리고 국내에서는 독자 AI 파운데이션 개발 프로젝트 추진되며 5개 기업들이 선정되어 자체 AI 기술력 확보를 적극 추진하고 있다.

이와 같이 국내외 AI 기업들은 기술력 확보와 생태계 영향력 확대를 위해 적극 노력하고 있으며 그 변화 속도는 매우 빠르다. 따라서 본 연구는 국내외 기업들의 동향을 조사를 통해 오픈소스AI가 기업과 AI 산업에 어떻게 영향을 주는지 파악하고자 한다.

따라서, 본 연구의 동향 조사는 기업 공식 발표, 논문, 보고서, 뉴스, 허깅페이스 등 다양한 자료 수집을 통해 수행하였다. 특히 각 기업들의 대표 모

델의 정량 데이터는 허깅페이스, 깃허브, 중국의 ModelScope 등의 오픈소스 협업 플랫폼의 자료들도 활용하였다. 그리고 국내 기업 자료는 상대적으로 많지 않아 개별 기업 자료 및 뉴스 같은 언론 자료들을 주로 활용하였다.

그리고 오픈소스AI 기업의 주요 조사 대상으로는 해외 아래 표와 같이 선정하였다. 해외 기업들은 오픈소스AI 생태계에서 영향력을 발휘하는 대표 미국 기업인 메타, 구글, OpenAI와 최근 부상하는 중국 기업인 알리바바, 바이두, 딥시크와 유럽의 대표 AI 기업인 미스트랄AI를 선정하였다. 그리고 국내 기업들은 독자 AI 파운데이터 모델 개발 프로젝트를 추진하고 있는 네이버, LG AI, SK 텔레콤, 업스테이지, NC AI를 선정하여 이들 기업들의 주요 동향과 대표 모델들의 현황을 조사하였다.

<표 3> 국내외 주요 오픈소스AI 기업

순번	국가	기업	주요모델	선정이유
1	미국	메타	Llama	오픈소스 LLM 점유율 45%
2		구글	Gemma, BERT, Gemini	다운로드 7.7억, GCP 매출 29% 성장 연계
3		오픈AI	GPT 시리즈	매출 $12.7B (2025 예상), 시장 선도
4		EleutherAI	GPT-Neo, GPT-J, Pythia	ThePile 데이터셋, 커뮤니티 오픈소스 대표
5	중국	알리바바	Qwen	9만 기업 고객, AI 클라우드 35.8% 점유
6		바이두	Ernie	MAU 3억, AI Cloud 42% YoY 성장
7		딥시크	DeepSeek-R1	$5.6M 혁신, Intelligence Score 68
8	유럽 (프랑스)	미스트랄 AI	Mistral	$14B 기업가치, 유럽 시장 40%
9	국내	네이버	HyperCLOVA X, SEED	검색 AI 국내 60%, 매출 $7.5B, 논문 공개 다수
10		LG AI	EXAONE	ThinQ AI 통합, $100M+ 비용 절감, 허깅페이스 20만
11		SK 텔레콤	A.X	T전화(에이닷) 1,000만 사용자, AI 매출 19% 성장
12		업스테이지	Solar	허깅페이스 50만, 300개 B2B 고객
13		NC AI	VARCO	게임/크리에이티브 특화

(자체 작성)

제2절 글로벌 오픈소스AI 기업 현황

1. 메타

메타는 오픈소스와 자사 플랫폼 통합을 결합한 듀얼 트랙 전략을 구사하고 있다. 라마는 WhatsApp, 인스타그램, 페이스북, 메신저에 통합되어 전 세계 30억 사용자에게 AI 어시스턴트 서비스를 제공한다. 이는 사용자 데이터 수집과 모델 개선을 위한 대규모 실험장(testbed) 역할을 한다.

엔터프라이즈 시장에서는 Accenture, AT&T, BNP Paribas 등이 Llama를 도입해 실질적 성과를 내고 있다. Accenture는 ESG 리포팅에 Llama 3.1을 적용해 생산성을 70% 향상시켰고, AT&T는 고객 응대 정확도를 33% 개선[24]하였다. BNP Paribas는 내부 LLM-as-a-Service를 구축[25]해 데이터 주권 확보와 업무 효율화를 동시에 달성하였다.

신사업 영역에서는 AR/VR 하드웨어와 AI의 융합을 추진한다. 보급형 VR인 Quest 3S($299)[26]는 Llama 기반 AI 기능을 통해 실시간 음성·제스처 인식과 공간 이해 능력을 강화하였다. 또한 Samsung과의 전략적 파트너십을 통해, Galaxy 기기 생태계에 Llama 최적화 모델 적용을 논의하며 모바일 AI 점유율 확대를 꾀하고 있다(Samsung은 2025년까지 4억 대 기기에 AI 적용 목표)[27]

수익 모델은 직접적인 라이선스 수익이 아닌 ① 자사 플랫폼 사용자 증대, ② 클라우드 파트너십(AWS, Azure, GCP와 협력), ③ 하드웨어 판매(Llama 최적화 칩셋)로 구성된다. 메타 Reality Labs는 Llama 기반 AR/VR 어시스턴트를 2025년 하반기 출시[28] 예정이다.

---

24) Quantum Zeitgeist, 2025, "Llama AI Model Sees Widespread Adoption Across Industries (Accenture, AT&T cases)", https://quantumzeitgeist.com/llama-ai-model-sees-widespread-adoption-across-industries

25) BNP Paribas, 2025, "BNP Paribas provides its businesses with an LLM as a Service platform", https://group.bnpparibas/en/press-release/bnp-paribas-provides-its-businesses-with-an-llm-as-a-service-platform-to-accelerate-the-deployment-of-generative-a

26) CNBC, 2024, "Meta unveils $299 Quest 3S VR headset", https://www.cnbc.com/2024/09/25/메타-unveils-cheaper-299-quest-3s-vr-headset-.html

27) AI Magazine, 2025, "Samsung's Galaxy AI Expansion: Aiming for 400m Devices (Mainly Google Gemini)", https://aimagazine.com/news/inside-samsungs-plans-to-bring-galaxy-ai-to-400m-devices, Tech Giants Pursue AI Partnerships with Samsunghttps://www.aiplusinfo.com/tech-giants-pursue-ai-partnerships-with-samsung/

28) Meta, 2025, "Meta Connect 2025 & Our Inaugural LlamaCon (September 17-18)", https://www.meta.com/blog/connect-2025-llamacon-save-the-date/,

메타의 2024년 총 매출은 $164.5B로 전년($134.9B) 대비 22% 성장했으며, AI 기반 광고 최적화가 이를 견인[29]하였다.

오픈소스 생태계에서는 Llama가 다운로드 10억 회를 돌파하며 약 45%의 점유율로 시장을 장악하였다. 2025년에는 AI 인프라 확장을 위해 자본 지출(CapEx)을 $60B~$65B 규모로 대폭 늘릴 계획[30]이며, Reality Labs(AR/VR)에 대한 지속적 투자와 함께 생성형 AI 및 GPU 클러스터 구축에 자원을 집중할 방침이다.

2. 구글

구글은 젬마(Gemma)를 아파치 2.0 라이선스로 완전 오픈소스화하여[31] 외부 기업들의 자유로운 상업 이용을 허용한다. 주요 기업 적용 사례로는 Shopify(Gemma 2 27B로 상품 추천 시스템 구축)[32], Replit(CodeGemma로 코드 자동 완성)[33] 등이 있다. Gemma의 허깅페이스 다운로드 7.7억 회는 Google이 오픈소스 AI 생태계에서 강력한 입지를 확보했음을 보여준다.

반면에 삼성과의 파트너십[34]을 통해서는 스마트폰 갤럭시 기기 4억 대에 Gemini Nano 및 Gemini Pro 기반의 AI 기능을 탑재하는 것을 목표로 하며, 이는 온디바이스 AI 시장을 장악을 추진하고 있다.

나아가 Gemini 시리즈(클로즈드 모델)를 자사 제품 전반에 통합하여 경쟁력을 강화하고 있다. 구글 검색 엔진은 Gemini Pro 기반의 AI Overviews 기능을 제공하여 2025년 기준 월간 20억 명[35] 이상의 사용자에게 도달[36]했으

---

29) 메타 Investor Relations, 2025, "Meta Reports Fourth Quarter and Full Year 2024 Results", https://investor.atmeta.com/investor-news/press-release-details/2025/Meta-Reports-Fourth-Quarter-and-Full-Year-2024-Results

30) Captide, 2025, "Meta Q4 2024 Earnings Analysis", https://www.captide.ai/insights/Meta-q4-2024

31) MarkTechPost, 2025, "Gemma 3 License Guide (Custom License, Not Apache 2.0)", https://markaicode.com/gemma-3-apache-license-commercial-use-guide/

32) Digital Commerce 360, 2025, "Shopify partners with Liquid AI (outperforming Gemma/Qwen)", https://www.digitalcommerce360.com/2025/11/13/shopify-liquid-ai-product-search-recommendations/

33) Google Cloud Case Study, 2025, "Replit case study: Powering Replit Agent with Gemini and Claude on Vertex AI", https://cloud.google.com/customers/repli

34) Samsung Newsroom, 2025, "Samsung and Google Cloud Join Forces To Bring Generative AI to Galaxy S24 (Gemini Pro/Imagen 2)", https://news.samsung.com/global/samsung-and-google-cloud-join-forces-to-bring-generative-ai-to-samsung-galaxy-s24-series

35) SQ Magazine, 2025, "ChatGPT vs. Google Gemini Statistics 2025 (AI Overviews reach 2B users)", https://sqmagazine.co.uk/chatgpt-vs-google-gemini-statistics/

36) Search Engine Journal, 2025, "Google Claims AI Overviews Monetize At Same Rate As Traditional Search", https://www.searchenginejournal.com/google-claims-ai-overviews-monetize-at-same-rate-as-traditional-search/547838/

며, 기존 검색과 대등한 광고 수익 효율을 달성하였다. 안드로이드 OS에는 Gemini Nano가 탑재되어 2025년까지 30억 대 기기에서 온디바이스 AI 기능을 제공할 예정[37]이다. 구글 클라우드(Vertex AI)는 엔터프라이즈 고객의 Gemini API 도입이 급증하며 Fortune 500 기업의 60% 이상[38]을 고객으로 확보하였다.

알파벳의 2024년 총 매출은 $350B이며, 이 중 구글 클라우드는 약 $42B (전년 대비 29% 이상 성장)를 기록하며 성장세를 주도하였다. AI 관련 매출은 Vertex AI의 엔터프라이즈 도입 가속화(사용량 전년 대비 20배 증가)에 힘입어 클라우드 부문 수익성에 크게 기여하고 있다.

오픈소스 생태계에서 젬마 시리즈는 1.5억 회 이상의 다운로드[39]를 기록하며 글로벌 입지를 다졌고, EmbeddingGemma는 500M 이하 소형 모델 중 벤치마크 1위를 달성하며 효율성을 입증하였다.

3. OpenAI

OpenAI는 2019년 GPT-2 출시 이후 모델의 가중치와 학습 코드를 비공개하는 '클로즈드 API(Closed API)' 전략으로 전환하여 기술 보안과 수익성을 동시에 확보하였다. 이러한 기조는 GPT-4o(2024)에 이어 2025년 공개된 o3-mini 및 o3 등 최신 모델에도 동일하게 적용되어 모든 주요 모델은 API 형태로만 제공[40]되고 있다.

외부 기업들은 이 API를 활용해 자사 제품에 생성형 AI 기능을 통합하고 있으며, Microsoft Azure OpenAI Service는 엔터프라이즈 확산의 핵심 통로 역할을 수행한다. 대표적으로 Morgan Stanley[41]는 GPT-4 기반으로 10만 건 이상의 리서치 문서를 분석하는 AI 어시스턴트를 운용 중이며, Stripe는 금융 사기 탐지 및 거래 패턴 분석을 고도화하였다. 또한 Duolingo와 Khan Academy는 각각 맞춤형 언어 교육과 AI 튜터 'Khanmigo'를 통해 개인화된

37) ReelMind AI, 2025, "Gemini Nano On Device: AI's Mobile Technology", https://reelmind.ai/blog/gemini-nano-on-device-ai-s-mobile-technology
38) Google Cloud, 2025, "Gen AI Unicorns and Fortune 500 Adoption", https://cloud.google.com/customers
39) TechCrunch, 2025, "Google's Gemma AI models surpass 150M downloads", https://techcrunch.com/2025/05/12/googles-gemma-ai-models-surpass-150m-downloads/
40) TechTarget, 2025, "OpenAI o3 and o4 explained: Everything you need to know", https://www.techtarget.com/whatis/feature/OpenAI-o3-explained-Everything-you-need-to-know
41) OpenAI, 2024, "Morgan Stanley uses AI evals to shape the future of financial services", https://openai.com/index/morgan-stanley/

교육 서비스[42]를 구현했으며, Shopify[43]는 고객 지원 자동화 및 마케팅 카피 생성에 해당 기술을 적극 활용하고 있다.

B2C 영역에서 OpenAI는 소비자 대상 서비스인 ChatGPT를 핵심 '캐시카우'로 성장시켰는데, 2025년 10월 기준 ChatGPT의 주간 활성 사용자(WAU)는 8억 명[44]을 돌파하며 역사상 가장 빠르게 성장한 소비자 애플리케이션으로 기록되었다.

수익 모델 또한 사용자 니즈에 맞춰 정교하게 세분화[45]되었다. 제한적 무료 사용 외에도 개인용 고급 기능을 제공하는 Plus($20/월), 소규모 협업을 위한 Team($25/월), o1 등 추론 모델의 무제한 활용을 지원하는 전문가용 Pro($200/월), 그리고 엔터프라이즈급 보안을 갖춘 Enterprise(맞춤형) 요금제로 구성된다. 이러한 다각화된 모델을 통해 2025년 하반기 기준 유료 구독자는 1,100만 명[46]을 넘어섰으며, 기업용 시트 사용자 역시 150만 명 이상을 확보하였다.

나아가 OpenAI는 단순 텍스트 생성을 넘어 멀티모달 및 에이전트 생태계를 완성하는 신사업 확장에 주력하고 있다. 300만 개 이상의 맞춤형 챗봇이 등록된 GPT Store는 크리에이터 수익 분배 시스템을 통해 자생적 생태계를 구축했으며, 2025년 9월 정식 출시된 비디오 생성 AI인 Sora 2[47]는 헐리우드 및 광고 산업의 영상 제작 파이프라인에 빠르게 통합되고 있다. 이 외에도 자율 실행형 AI 에이전트인 Operator와 고정밀 음성 인식 및 이미지 생성 기능을 제공하는 Whisper, DALL-E 3 등을 통해 전방위적인 AI 솔루션 포트폴리오[48]를 완성해 나가고 있다.

OpenAI의 2024년 총 매출은 $3.7B로, 2023년 대비 4배 성장하였다. 2025

42) Forbes, 2024, "OpenAI Launches GPT Store: Where Creators Can Share—And Possibly Make Money", https://www.forbes.com/sites/mollybohannon/2024/01/10/openai-launches-gpt-store-where-creators-can-share-and-possibly-make-money-from-their-chatbots/
43) MESA, 2025, "What is Shopify Sidekick? AI Agent for Merchants in 2025", https://www.getmesa.com/blog/shopify-sidekick/
44) TechCrunch, 2025, "Sam Altman says ChatGPT has hit 800M weekly active users", https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users
45) OpenAI, "ChatGPT Pricing", https://openai.com/chatgpt/pricing/
46) The Information, 2024, "OpenAI COO Says ChatGPT Passed 11 Million Paying Subscribers", https://www.theinformation.com/articles/openai-coo-says-chatgpt-passed-11-million-paying-subscribers
47) OpenAI, 2025, "Sora 2 is here", https://openai.com/index/sora-2/
48) Forbes, 2024, "OpenAI Launches GPT Store: Where Creators Can Share—And Possibly Make Money", https://www.forbes.com/sites/mollybohannon/2024/01/10/openai-launches-gpt-store-where-creators-can-share-and-possibly-make-money-from-their-chatbots

년에는 $12.7B~$13B 매출을 예상하고 있으며, 이는 ChatGPT 유료 구독자 증가(1,000만 명 이상)와 API 사용량 급증(월간 100억 요청 이상)에 기인[49] 한다. 글로벌 생성형 AI 시장에서 OpenAI는 약 34%의 시장 점유율을 차지하며 압도적 1위를 유지하고 있다.

4. Eleuther AI

EleutherAI는 2020년 GPT-3의 비공개에 반발하여 설립된 비영리 연구 컨소시엄으로, “누구나 접근 가능한 대규모 언어 모델 개발“이라는 핵심 철학을 바탕으로 운영된다. 수익 모델이 없는 순수 비영리 조직임에도 불구하고 CoreWeave(GPU 클러스터), Stability AI(자금), Hugging Face(인프라), Canva 등의 여러 기업 후원을 통해 지속 가능한 연구 환경을 구축하였다. 외부 기업 및 연구기관들은 EleutherAI가 공개한 모델과 도구를 자유롭게 활용하고 있으며, Stanford HAI는 2025년 보고서[50]에서 이를 “오픈소스 AI 운동의 상징“으로 평가하였다.

EleutherAI의 주요 기여[51]는 AI 연구의 투명성과 재현성을 획기적으로 높인 세 가지 핵심 자산으로 요약된다. 첫째, The Pile 데이터셋은 825GB 규모의 고품질 텍스트 데이터로, 22개 도메인을 아우르며 학계와 산업계에서 표준 학습 데이터로 널리 사용되고 있다. 둘째, Pythia Suite[52]는 70M에서 12B 파라미터 모델을 학습 단계별로 저장한 154개 체크포인트를 공개하여, 블랙박스로 여겨지던 LLM의 학습 과정을 투명하게 분석할 수 있게 하였다. 셋째, Evaluation Harness[53]는 현재 200개 이상의 벤치마크를 지원하는 언어 모델 평가의 표준 프레임워크로, 전 세계 연구자들이 모델 성능을 객관적으로 비교・검증하는 데 필수적인 도구로 자리 잡았다.

EleutherAI는 커뮤니티 기반의 분산 협력 모델이라는 독창적인 운영 방식을 채택하고 있다. Discord와 GitHub를 중심으로 전 세계 100~200명의 활동 기여자가 자발적으로 연구에 참여하며, 이러한 집단 지성은 거대 자본 없이

49) CNBC, 2025, "OpenAI hits $10 billion in annualized revenue", https://finance.yahoo.com/news/openai-breaks-10-billion-arr-010416720.html
50) Stanford HAI, 2025, "The 2025 AI Index Report", https://hai.stanford.edu/ai-index/2025-ai-index-report
51) TechCrunch, 2023, "Stability AI, Hugging Face and Canva back new AI research nonprofit", https://techcrunch.com/2023/03/02/stability-ai-hugging-face-and-canva-back-new-ai-research-nonprofit/
52) Hugging Face, "EleutherAI Pythia Suite", https://huggingface.co/EleutherA
53) GitHub, 2025, "LM Evaluation Harness", https://github.com/EleutherAI/lm-evaluation-harness

도 혁신이 가능함을 증명하였다. 특히 대표작인 GPT-Neo와 GPT-J는 자원 봉사자들의 분산 협력[54]만으로 개발된 대규모 언어 모델로, AI 개발의 민주화를 실증한 역사적 사례로 꼽힌다.

기술 개발뿐만 아니라 AI 윤리와 안전성 연구에서도 선도적인 역할을 수행하고 있다. 커뮤니티는 AI 정렬 포럼(AI Alignment Forum) 및 레드팀 활동에 적극 참여하며, 잠재적 위험을 선제적으로 식별하고 있다. 또한 AI 정책 및 규제에 관한 심도 있는 연구 보고서를 지속적으로 발표하여, 학계와 정책 입안자들에게 기술적 통찰과 규제 가이드라인을 제공하는 중요한 싱크탱크 역할을 수행하고 있다.

5. 알리바바

알리바바는 전세계 4위의 클라우드 서비스 기업이자 아시아 최대 클라우드 기업이다. 중국 AI 클라우드 서비스 시장에서 35.8%[55]의 점유율로 압도적 1위를 차지하고 있으며, 글로벌 오픈소스AI 생태계에서도 Qwen 시리즈는 2025년 6월 기준 다운로드 2,000만 회 이상(Hugging Face/ModelScope 합산)을 기록하며 영향력을 확대하고 있다.

알리바바 Qwen은 중국 최대 규모인 90,000개 이상[56]의 기업 고객 기반을 보유한 오픈소스 AI 플랫폼으로, Apache 2.0 라이선스를 통해 글로벌 시장 공략에 적극적이다. Snowflake(데이터 분석), BNP Paribas(금융 컴플라이언스), Cisco(네트워크 자동화) 등 유수의 글로벌 기업들이 이를 도입해 실질적인 성과를 거두고 있으며, 일본과 한국의 스타트업들 또한 Qwen 기반 서비스 개발을 통해 비용 효율성을 높이고 있다.

Alibaba는 인프라부터 응용 서비스까지 아우르는 '풀스택(Full-Stack) AI 전략'을 구사한다. 자체 개발한 Hanguang 800 NPU는 500 IPS/W의 뛰어난 전력 효율로 AI 추론 비용을 50% 이상 절감[57]시키는 핵심 동력이며, 이를 30

54) EleutherAI, 2025, "About EleutherAI", https://www.eleuther.ai/about
55) SCMP, 2025, "Alibaba holds wide lead over rivals in China's AI cloud market (35.8%)", https://www.scmp.com/tech/big-tech/article/3325034/alibaba-holds-wide-lead-over-rivals-bytedance-huawei-tencent-chinas-ai-cloud-market
56) Alibaba Cloud, 2024, "Qwen LLM Tops 90,000 Enterprise Clients", https://www.alibabacloud.com/blog/alibaba-clouds-qwen-models-attract-over-90000-enterprise-adoptions-within-its-first-year_601133
57) Network World, 2025, "Alibaba is developing an AI inference chip amid US export curbs", https://www.networkworld.com/article/4049120/alibaba-is-developing-an-ai-inference-chip-amid-us-export-curbs.htm

개 리전의 Alibaba Cloud 인프라와 결합해 압도적인 가격 경쟁력을 제공한다. 또한 중국판 Hugging Face인 ModelScope 플랫폼을 통해 오픈소스 생태계를 장악하며 Qwen 모델 확산을 가속화하고 있다.

이러한 기술력은 Taobao, Tmall, DingTalk[58] 등 자사 플랫폼에 깊숙이 통합되어 10억 명 이상의 사용자에게 AI 서비스를 제공하고 있다. Taobao와 Tmall은 큐웬 기반 AI 추천 시스템을 도입해 클릭률(CTR)과 구매 전환율을 10~25% 향상[59]시켰으며, 판매자에게 개인화된 상품 설명 자동 생성 기능을 제공해 운영 비용을 30% 절감[60]시켰다. DingTalk 협업 툴 역시 Qwen을 통합하여 회의 요약, 문서 자동 작성, 스마트 검색 기능을 7억 명의 기업 사용자[61]에게 제공하며 업무 생산성을 혁신하였다.

수익 모델은 알리바바 클라우드 API($0.50/백만 토큰), 엔터프라이즈 라이선스, 커스텀 모델 학습 서비스 등으로 다각화되어 있다. 알리바바 클라우드는 중국 AI 클라우드 시장에서 35.8%의 점유율[62]로 압도적 1위를 유지하고 있으며, 2024년 말 기준 클라우드 매출은 AI 수요 폭증에 힘입어 전년 대비 34% 급증하였다. 더불어 신사업 영역에서는 Qwen 기반의 멀티모달 인식 기술을 활용한 자율주행, 중국 공장 자동화를 이끄는 스마트 제조, 그리고 K-12 맞춤형 학습을 지원하는 교육 AI 분야로 확장을 가속화[63]하며 중국 정부의 디지털 전환 정책과 긴밀히 연계된 성장 전략을 펼치고 있다.

알리바바의 2024 회계연도 총 매출은 $130.4B이며, 이 중 알리바바 클라우드는 AI 수요 폭증에 힘입어 2025년 상반기(4~9월) 전년 대비 26~34%[64]의 폭발적인 매출 성장을 기록하였다. 특히 AI 관련 제품 매출은 8분기 연속 세 자릿수 성장률을 보이며 전체 공용 클라우드 매출의 20% 이상을 차지하게 되었다.

---

58) Alizila, 2025, "Alibaba's DingTalk Hits 700M Users and Launches AI Agent", https://www.alizila.com/alibaba-dingtalk-700m-users-2023-ai-agent-boost-workspace-productivity

59) Complete AI Training, 2025, "Alibaba deploys Qwen AI across Taobao and Tmall for double-digit lifts", https://completeaitraining.com/news/alibaba-deploys-qwen-ai-across-taobao-and-tmall-for-double

60) Futu News, 2025, "This Tmall Double 11, AI has become the 'operator' of 5 million merchants", https://news.futunn.com/en/post/64429252/a-full-store-inspection-in-3-minutes-with-materials-automatically

61) AIBase, 2025, "AliQwen APP Launches Qwen3-Learning Large Model", https://www.aibase.com/news/2339

62) CNBC, 2025, "Alibaba shares rise as AI drives 34% cloud sales jump", https://www.cnbc.com/2025/11/25/alibaba-shares-rise-as-ai-drives-cloud-sales-jump-earnings.html

63) ConnectCX, 2025, "Alibaba Cloud's Qwen: Powering China's AI-Driven Industrial Revolution", https://connectcx.ai/alibaba-clouds-qwen-powering-chinas-ai-driven-industrial-revolution

64) CNBC, 2025, "Alibaba shares rise as AI drives 34% cloud sales jump", https://www.cnbc.com/2025/11/25/alibaba-shares-rise-as-ai-drives-cloud-sales-jump-earnings.html

R&D 투자는 연간 $15B 규모를 유지하되, 향후 3년간 AI 인프라 및 데이터센터 구축에 $53B(약 3,800억 위안) 이상을 집중 투자한다는 'AI 인프라 대전환' 계획[65]을 2025년 2월 발표하였다.

6. 바이두

바이두는 2025년 6월 30일, 전략적 대전환을 통해 최신 모델인 Ernie 4.5[66]를 Apache 2.0 라이선스로 완전 오픈소스화하였다. 이는 DeepSeek의 성공에 자극받아 폐쇄형 생태계에서 개방형 생태계로 선회한 결정으로, 중국 정부 기관은 물론 ICBC, China Construction Bank 등 금융권과 K-12 교육 플랫폼, 대형 병원 등 공공 및 민간 부문의 폭넓은 도입을 이끌어내고 있다.

알리바바와 유사한 풀스택 AI 전략을 구사하는 바이두는 자체 개발한 모델(Ernie), 칩셋(Kunlun), 프레임워크(PaddlePaddle), 클라우드(Baidu Cloud)를 수직 계열화하여 운영 효율을 극대화하였다. 특히 3만 개의 Kunlun P800 NPU[67]로 구성된 단일 클러스터를 가동하여 AI 컴퓨팅 자립을 실현했으며, 이를 기반으로 Baidu Cloud는 고성능 모델인 'Ernie 4.5 Turbo'를 백만 토큰당 $0.15라는 파격적인 가격에 제공[68]하고 경량 모델인 'Ernie Lite'는 무료로 개방하여 가격 경쟁력을 확보하였다.

이러한 AI 기술력은 검색 엔진, 지도, 클라우드, 자율주행 등 기존 핵심 사업 전반에 통합되어 경쟁력을 강화하고 있다. 월간 6억 명 이상의 사용자가 활용하는 바이두 검색 엔진에는 Ernie Bot이 통합되어 AI 기반의 정교한 검색 결과를 제공하며, 바이두 Maps는 자연어 내비게이션과 실시간 교통 예측 기능을 통해 중국 지도 시장 점유율 1위를 굳건히 지키고 있다. 또한 Baidu Cloud의 AI Cloud 사업 부문은 2025년 1분기 기준 전년 동기 대비 42%[69] 성장한 67억 위안($940M)의 매출을 기록하며 그룹의 새로운 성장 엔

65) Bloomberg, 2025, "Alibaba Plans to Spend $53 Billion on AI Infrastructure", https://www.bloomberg.com/news/articles/2025-02-24/alibaba-to-spend-53-billion-on-ai-infrastructure-in-big-pivo

66) TechTarget, 2025, "Baidu makes foundation model Ernie 4.5 open source", https://www.techtarget.com/searchenterpriseai/news/366626838/Baidu-makes-foundation-model-Ernie-45-open-source

67) TurtlesAI, 2025, "Baidu powers up Kunlun super-cluster (30,000 P800 chips)", https://www.turtlesai.com/en/pages-2707/baidu_powers_up_kunlun_super_cluster_unveils_cutti

68) PR Newswire, 2025, "Baidu Launches ERNIE 4.5 Turbo with competitive pricing", https://www.prnewswire.com/news-releases/baidu-launches-ernie-4-5-turbo-ernie-x1-turbo-and-new-suite-of-ai-tools-to-empower-deve-302126245.htm

진으로 자리 잡았다.

Ernie가 적용된 신사업 영역 중 가장 주목받는 분야는 자율주행 플랫폼 Apollo Go다. Apollo Go[70]는 중국 내 Wuhan, Shenzhen, Beijing 등 10개 도시에서 로보택시 서비스를 제공하며 누적 1억 회 이상의 주행 기록을 달성하였다. 나아가 2025년 7월 Uber, 8월 Lyft와 연이어 전략적 파트너십을 체결하며 글로벌 시장 진출을 본격화했으며, 2025년 말까지 1,000대 이상의 완전 자율주행 차량을 운영할 계획이다.

바이두의 2024년 총 매출은 약 $19B(약 1,340억 위안)이며, 이 중 AI 관련 매출 비중이 급격히 확대되고 있다. 2025년 1분기 AI Cloud 매출은 전년 동기 대비 42%[71] 급증한 $940M(약 67억 위안)을 기록하였다. 이는 Ernie API의 엔터프라이즈 도입 가속화와 생성형 AI 워크로드 증가에 기인한다.

중국 AI 챗봇 시장에서 Ernie Bot은 2024년 6월 기준 사용자 수 3억 명을 돌파하며 선두를 유지하고 있다. 경쟁 서비스인 ByteDance의 Doubao, 알리바바의 Tongyi와 치열한 점유율 경쟁 중이다. 중국 AI 클라우드 시장에서는 6~15%의 점유율(기관별 상이)을 차지하며 알리바바, ByteDance, 화웨이 등과 함께 'AI 4강' 구도를 형성[72]하고 있다.

7. 딥시크

딥시크는 모든 모델을 MIT 라이선스로 완전 오픈소스화하여 외부 기업들의 자유로운 상업 이용을 허용하는 파격적인 전략을 취하고 있다. 2025년 AI 업계의 최대 이슈 메이커로 떠오른 DeepSeek-R1은 OpenAI o1과 대등한 추론 성능을 보이면서도, 학습 비용은 불과 560만 달러[73]에 그쳐 OpenAI GPT-4($100M+) 대비 약 1/18 수준의 혁신적인 비용 효율성을 입증하였다. 이는 "Jevons Paradox(효율성이 증가하면 수요가 폭발하는 현상)"를 촉발하

69) GuruFocus, 2025, "Q1 2025 Baidu Inc Earnings Call Transcript (AI Cloud +42%)", https://www.gurufocus.com/stock/BIDU/transcripts/2881896

70) Gasgoo, 2025, "Baidu's Apollo Go fast-tracks global push with Uber, Lyft alliances", https://autonews.gasgoo.com/m/70038509.html

71) EarningsIQ, 2025, "Baidu Q1 2025: AI Cloud Jumps 42%, Reshaping Core Revenue Mix", https://www.earningsiq.co/articles/baidu_bidu_q1_2025_ai_cloud_jumps_42_reshaping_core_revenue

72) SCMP, 2025, "Alibaba holds wide lead over rivals (Baidu 6.1% share)", https://www.scmp.com/tech/big-tech/article/3325034/alibaba-holds-wide-lead-over-rivals-bytedance-huawei-tencent-chinas-ai-cloud-marke

73) RD World, 2025, "DeepSeek-R1 RL model: 95% cost cut vs. OpenAI's o1 ($5.6M training cost)", https://www.rdworldonline.com/this-week-in-ai-research-a-0-55-m-token-model-rivals-openais-60-flagship/

며 전 세계적으로 AI 인프라 수요를 급증시키는 계기가 되었다.

주요 기업 고객으로는 ByteDance(Doubao AI 서비스에 딥스크 통합[74]), 텐센트(WeChat AI 기능 강화)[75], 샤오미(스마트폰 온디바이스 AI 탑재)[76] 등 중국 빅테크들이 대규모 도입을 진행 중이다. 특히 딥시크는 PTX 프로그래밍을 통해 NVIDIA CUDA에 대한 의존성을 최소화[77]하고 Huawei Ascend, Biren 등 중국 국산 AI 칩과의 호환성[78]을 확보하여 중국 AI 주권 전략의 핵심 요소로 자리 잡았다. 또한 NVIDIA RTX 4090 등 소비자급 GPU에서도 실행 가능한 온디바이스 배포를 지원하여 접근성을 극대화하였다.

핵심 사업인 딥시크 API 플랫폼은 '가격 파괴'를 통해 시장 점유율을 빠르게 확대하고 있다. 입력 토큰(Cache Hit) 기준 가격은 $0.14/백만 토큰[79]으로, OpenAI o1($15.00/백만 토큰) 대비 약 1/100 수준이라는 압도적인 가격 경쟁력을 제공한다. 자체 챗봇 서비스인 DeepSeek Chat은 2025년 3월 기준 월간 활성 사용자(MAU) 3,300만 명을 돌파[80]하며 중국 내 무료 AI 챗봇 시장에서 가장 빠른 성장세를 보이고 있다.

딥시크의 모기업인 High-Flyer Capital Management(양적 헤지펀드)는 금융 AI 알고리즘 개발에 딥시크 모델을 적극 활용하고 있다. 주식 거래 전략 수립, 리스크 관리, 시장 예측 등에 AI를 적용하여 운용 성과를 고도화하고 있으며, 이는 AI 기술이 실제 금융 수익 창출에 기여하는 대표적인 사례로 꼽힌다. Stanford HAI는 2025년 AI 인덱스 보고서에서 DeepSeek를 "2025년 가장 영향력 있는 AI 기업" 중 하나로 선정하였다.

DeepSeek은 비상장 스타트업으로 구체적인 재무 데이터를 공개하지 않지만, 2025년 기준 기업 가치는 $3.4B(약 4.7조 원) 이상으로 추정된다. API 서비스 수익은 2025년 $50M~$100M 규모로 예상[81]되나, 당장의 수익성보다는

---

74) TechNode, 2025, "Tesla taps ByteDance, DeepSeek to power AI in China", https://technode.com/2025/08/26/tesla-taps-bytedance-deepseek-to-power-ai-in-china/

75) China Daily, 2025, "WeChat embracing DeepSeek for tech leap (AI Search integration)", https://www.chinadailyhk.com/hk/article/604828

76) XiaomiTime, 2025, "Xiaomi announced official DeepSeek R1 supported devices (HyperOS integration)", https://xiaomitime.com/xiaomi-announced-official-deepseek-r1-supported-devices-23865/

77) Tom's Hardware, 2025, "DeepSeek's new AI model debuts with support for China-native chips (CANN support)", https://www.tomshardware.com/tech-industry/deepseek-new-model-supports-huawei-cann

78) The Reach, 2025, "Huawei's New Chip Might Be The Companion DeepSeek Needed (PTX & Ascend)", https://thereach.ai/2025/02/05/huaweis-new-chip-might-be-the-companion-deepseek-needed/

79) DeepSeek R1 Pricing, 2025, "$0.14 vs OpenAI o1 $15.00 comparison", https://deepseek-r1.com/pricing

80) Thunderbit, 2025, "DeepSeek Statistics 2025: 33.7M MAU & $5.6M training cost", https://thunderbit.com/blog/deepseek-ai-statistics

81) Miracuves, 2025, "DeepSeek Revenue Model 2025 ($1.1B Revenue Forecast)", https://miracuves.com/blog/deepseek-revenue-model

초저가 전략($0.14/백만 토큰)을 통해 사용자 기반을 확보하고 생태계를 장악하는 데 주력하고 있다. 글로벌 오픈소스 LLM 생태계에서는 허깅페이스 다운로드 수 1.2억 회를 기록[82]하며 상위권에 랭크되었고, 자체 벤치마크 점수(Intelligence Score)에서 글로벌 오픈소스 모델 중 1위를 차지[83]하며 기술력을 과시하였다.

딥시크의 등장은 글로벌 AI 산업에 '딥시크 쇼크'라 불리는 엄청난 파급효과를 미쳤다. 2025년 1월 27일, DeepSeek-R1 공개 직후 NVIDIA 주가는 17% 폭락하며 하루 만에 시가총액 $589B(약 820조 원)가 증발[84]하였다. 이는 AI 모델 개발이 더 이상 거대 자본을 가진 빅테크의 전유물이 아니며, 소규모 팀도 효율적인 아키텍처와 알고리즘 최적화만으로 충분히 경쟁할 수 있음을 입증한 역사적 사건으로 평가된다. 이에 따라 OpenAI 등 경쟁사들도 가격 인하 압박에 직면하며 글로벌 AI 시장의 가격 구조가 재편되고 있다.

8. 미스트랄 AI

미스트랄AI는 아파치 2.0 라이선스로 Mistral 7B, Mixtral 8x7B, Mixtral 8x22B 등을 오픈소스로 공개하여 타 기업들의 자유로운 상업 이용을 허용하고 있다. 주요 기업 고객은 BNP Paribas, Orange, Renault 등의 유럽 대표 기업들이다.

BNP Paribas[85]는 미스트랄AI와 전사적 파트너십을 체결하고 자체 구축한 'LLM as a Service' 플랫폼에 Mistral 모델을 직접 도입하였다. 이를 통해 데이터 보안을 유지하며 규제 문서 분석 및 컴플라이언스 업무를 자동화하고, 자회사인 Hello bank!의 챗봇 서비스 고도화에 활용하고 있다.

Orange[86]는 통신 네트워크 최적화 및 고객 서비스 자동화에 Mistral 모델

82) Reddit (LocalLLaMA), 2025, "Deepseek R1 just became the most liked model ever on Hugging Face", https://www.reddit.com/r/LocalLLaMA/comments/1ipz13t/deepseek_r1_just_became_the_most_liked_model_ever/
83) Artificial Intelligence News, 2025, https://www.artificialintelligence-news.com/news/deepseek-v3-0324-tops-non-reasoning-ai-models-open-source-first/
84) CNBC, 2025, "Nvidia stock plummets, loses record $589 billion as DeepSeek prompts questions over AI spending", https://finance.yahoo.com/news/nvidia-stock-plummets-loses-record-589-billion-as-deepseek-prompts-questions-over-ai-spending-135543685.html
85) BNP Paribas, 2024, "BNP Paribas and Mistral AI sign a partnership agreement", https://group.bnpparibas/en/press-release/bnp-paribas-and-mistral-ai-sign-a-partnership-agreement-covering-all-mistral-ai-models
86) Orange, 2025, "Orange and Mistral AI Forge Strategic AI Partnership",

을 적용하여 연간 수천만 유로 규모의 비용 효율화를 달성했으며, Renault[87]는 자동차 설계 시뮬레이션에 AI를 접목하여 핵심 부품 개발 기간을 기존 4년에서 2년으로 50% 단축시키는 성과를 거두었다.

미스트랄AI는 Microsoft Azure와 AWS Bedrock를 통해 모델을 배포하여 글로벌 클라우드 플랫폼을 통한 접근성을 확대하고 있다. 이 과정에서 아파치 2.0 라이선스의 오픈소스 모델(Mistral 7B/Mixtral)로 생태계를 확장하고, 고성능 프리미엄 모델(Mistral Large/Medium)은 API로 제공하여 직접적인 수익을 창출하는 이중 전략을 추진하고 있다.

이 전략은 유럽의 'AI 주권(AI Sovereignty)' 목표와 맞물려, 미국 및 중국 기술 의존도를 낮추려는 유럽 공공 및 기업들의 수요를 흡수하고 있다. 특히 EU AI Act 준수를 위해 설명 가능한 AI(Explainable AI) 기능을 강화하며, 규제에 민감한 기업들에게 차별화된 가치를 제공하고 있다..

핵심 사업은 Mistral API 플랫폼, 자체 개발 플랫폼인 La Plateforme, 그리고 엔터프라이즈 맞춤형 솔루션으로 구성되어 있다. Mistral API[88]는 월간 100억 건 이상의 요청을 처리하며 유럽을 중심으로 빠르게 성장하고 있으며, La Plateforme은 개발자들이 Mistral 모델을 미세 조정(Fine-tuning)하고 배포할 수 있는 통합 환경을 제공하여 2024년 기준 10,000명 이상의 개발자가 이용하고 있다.

미스트랄AI의 2024년 매출은 $30M(약 2800만 유로)을 기록하며 2023년 보다 성장하였으며, 2025년에는 더욱 성장하여 상반기에만 $100M을 돌파할 것으로 예상[89]되고 있다. 이는 Mistral API 사용량 증가와 유럽을 넘어선 글로벌 엔터프라이즈 고객 확대 덕분이다. 또한 기업 가치 역시 급상승하여 2024년 6월 시리즈B 단계에서 $6.2B(58억 유로)였던 평가액이 2025년 9월 ASML이 주도한 시리즈C[90] 단계에서 $14B(약 117억 유로)에 크게 증가하였다.

유럽AI 시장에서 미스트랄AI는 데이터 주권과 EU AI Act 준수 수요를 바

---

https://newsroom.orange.com/orange-and-mistral-ai-join-forces-to-accelerate-artificial-intelligence-development-in-europe

87) Narratize, 2024, "Renault reduces design time by 50% with AI", https://www.narratize.com/blogs/ai-is-slashing-product-development-times

88) Scribd, 2025, "Mistral AI La Plateforme & API Usage Statistics", https://www.scribd.com/document/872095733/Mistral-Complete-Research

89) AIFundingTracker, 2025, "Mistral AI Revenue Growth: From Zero to $100M+", https://aifundingtracker.com/mistral-ai-funding-unicorn-valuation/

90) WSJ, 2025, "Mistral AI Doubles Valuation to $14 Billion With ASML Investment", https://www.wsj.com/tech/ai/asml-to-invest-1-5-billion-in-french-startup-mistral-ai-0d5eb547

탕으로 약 40% 이상의 점유율을 보이며 압도적 1위[91]를 지키고 있다. 특히 프랑스, 독일, 영국 등 주요 국가의 금융, 통신, 제조 분야 기업들이 핵심 고객이 되고 있다. 글로벌 오픈소스 LLM 생태계 영향력도 막강하여 허깅페이스에서 누적 다운로드 수는 2.5억 회를 넘어섰으며, 이를 기반으로 1.5만 개 이상의 파생 모델이 개발[92]되어 라마와 함께 오픈소스 진영의 양대 산맥으로 인식되고 있다.

<표 4> 해외 주요 오픈소스AI 기업 현황 요약

구분	주요 오픈소스 AI	법인 유형	AI 관련 주요 사업	오픈소스 전략 목적	주요 파트너십
메타	Llama 3.1 (405B) Llama4Scout/Maverick	상장사	SNS AI 추천, 광고 최적화, VR/AR	OpenAI 견제, 생태계 확보 (개발자 락인)	Microsoft Azure
구글	BERT (11만+ 인용) Gemma2(27B)	상장사	검색, 광고, 클라우드, 온디바이스 AI	온디바이스 AI Android생태계	AWS, GCP AndroidOEM 클라우드고객
OpenAI	GPT-2, gpt-oss-120B	비영리 →영리 전환	AI 서비스, 기술 공급	영향력 확보	Microsoft (독점) $13B투자
Eleuther AI	GPT-Neo (2.7B) GPT-J-6B Pythia	비영리	오픈소스 협업 (커뮤니티)	기술 협업, 학술공헌	CoreWeave StabilityAI후원
딥시크	DeepSeek-V3 (671B) DeepSeek-R1(추론 모델)	스타트업	AI 서비스, 기술 공급	개방형 검증 (영향력 확보)	API 플랫폼 (독립 운영)
알리바바	Qwen2.5 (0.5B-72B) Qwen3 (235B, MoE)	상장사	클라우드 AI, 전자상거래	중국 점유율 확보, 글로벌확장	Alibaba Cloud 전자상거래통합
바이두	Ernie 4.5 ErnieSpeed	상장사	검색, 자율주행, 클라우드	시장 경쟁력 ErnieAPI보완	Baidu Cloud Apollo자율주행
미스트랄 AI	Mistral 7B Mixtral8x7B Codestral	스타트업	기술 공급	유럽 AI 주권 미국/중국견제	Microsoft Azure AWS 유럽기업

(자체 작성)

91) Sifted, 2024, "Mistral set to hit €30m in revenues", https://sifted.eu/articles/mistral-ai-llm-market

92) Hugging Face Blog, 2025, "Model statistics: Mistral ranks 2nd in downloads after 메타", https://huggingface.co/blog/lbourdois/huggingface-models-stat

제3절 국내 주요 오픈소스AI 기업 현황

1. LG AI

LG AI는 초거대 AI인 EXAONE(엑사원)을 중심으로 LG그룹 전반의 AI 전환을 주도하고 있다. 2024년 8월 공개된 EXAONE 3.0은 오픈소스로 개방되어 생태계를 확장하는 한편, 엔터프라이즈용 고성능 모델은 상업 라이선스로 제공되어 수익성을 확보하고 있다. 특히 2025년 공개된 EXAONE 4.0은 하이브리드 추론 능력을 갖춰 한국어 및 법률, 의료 등 전문 도메인 영역에서 GPT-4 수준의 성능을 입증[93]하였다.

자사 비즈니스 통합 전략에서도 가시적인 성과를 거두고 있다. LG전자는 '공감 지능(Affectionate Intelligence)'을 탑재한 AI 가전 라인업을 대폭 확대하여 2024년 H&A(생활가전) 본부 매출 33.2조 원($24B)[94]을 기록하였다. LG U+는 EXAONE 기술을 경량화한 자체 모델 iXi-GEN을 AI 콜센터(AICC)에 적용하여 월평균 상담 시간을 117만 분 단축[95]하고 응대 효율을 19% 개선하였다.

또한 LG CNS는 EXAONE 기반의 기업용 AI 플랫폼 'Gen AI Studio'를 통해 300개 이상의 외부 기업 프로젝트를 수행하며 B2B 시장을 적극적으로 공략하고 있다. 제조 분야에서는 LG화학이 AI를 신소재 개발 파트너로 활용하여 실험 데이터 분석 및 후보 물질 발굴 기간을 획기적으로 단축시키는 등 R&D 효율성에 기여하도록 확산중이다. 이러한 전사적 AI 도입을 통해 LG 그룹은 연간 1,000억 원($75M) 이상의 비용 절감 효과를 거두고 있으며, 2025년에는 멀티모달 API 서비스와 'ChatEXAONE'의 계열사 전면 도입을 통해 시너지를 극대화할 계획[96]이다

LG AI는 독립 법인이 아닌 LG 경영개발원 산하의 연구 조직으로 별도의 매출이나 영업이익을 공시하지 않으나, 그룹 차원의 전략적 지원 하에 연간

93) LG AI Research, 2025, "EXAONE 4.0: Unified Large Language Models", https://arxiv.org/html/2507.11407v1

94) Korea Times, 2025, "LG Electronics achieves record revenue in 2024 (H&A 33.2 trillion won)", https://www.koreatimes.co.kr/business/companies/20250108/south-koreas-lg-electronics-achieves-record-revenue-in-202

95) Nate News, 2025, "LG Uplus cuts customer call times by 1.17M minutes with AI advisor", https://news.nate.com/view/20250528n10912

96) Google Cloud, 2025, "LG AI Research Success Story", https://cloud.google.com/customers/intl/ko-kr/lgai?hl=ko-KR

약 $500M(약 6,500억 원) 규모의 R&D 예산을 운용하는 것으로 추정[97]된다. 이는 국내 AI 연구소 중 최상위권의 투자 규모로, 안정적인 연구 환경과 인프라 확보의 기반이 되고 있다.

국내 생성형 AI 시장에서는 범용 서비스 중심의 네이버(HyperCLOVA X)와 통신 기반의 SK 텔레콤(A.X)에 이어, '제조 및 산업 특화 AI'라는 차별화된 포지셔닝으로 약 25% 내외의 시장 영향력을 행사하고 있다. 특히 정부가 지정한 '국가대표 AI 기업(Sovereign AI)' 5개 사[98]에 포함되어 공공 및 국가 과제 수주 경쟁력을 확보하였다.

글로벌 오픈소스 생태계 지표인 Hugging Face에서는 2024년 8월 모델 공개 이후 단기간에 누적 다운로드 8만 회 이상을 기록하며, 국내 기업 중에서는 이례적으로 글로벌 개발자 커뮤니티에 성공적으로 안착하였다. 기술적 성능 지표인 KMMLU-Pro(한국어 다중 과제 이해 평가) 벤치마크에서는 EXAONE 최신 모델이 GPT-4를 넘어서는 최고점을 갱신[99]하며, 한국어 및 전문 지식 추론 분야에서 '기술적 해자(Moat)'를 구축했음을 수치로 입증하였다.

2. SK 텔레콤

SK 텔레콤은 'AI 피라미드 전략'을 기반으로 ①AI 인프라(AI DC, 반도체), ②AIX(핵심 사업의 AI 전환), ③AI 서비스(A.dot 글로벌 확장)라는 3대 축을 구축하여 AI 수익화를 본격화하고 있다. 2025년에는 이를 한 단계 발전시킨 'AI 피라미드 2.0'[100]을 통해 단순 인프라 구축을 넘어 실질적인 수익 창출 구조를 완성해 나가고 있다.

핵심이 되는 자체 LLM 'A.X(에이닷엑스) 4.0'[101]은 오픈소스 진영에서 최

97) Pulse News, 2024, "LG Group encourages R&D investment", https://pulse.mk.co.kr/news/english/11151986
98) Korea Herald, 2025, "LG, SK 텔레콤, 네이버 selected for Korea's sovereign AI push", https://www.koreaherald.com/article/10546363
99) arXiv, 2025, "From KMMLU-Redux to Pro: A Professional Korean Benchmark", https://arxiv.org/html/2507.08924v1
100) Khan Academy (Khan News), 2025, "SK telecom 'AI Pyramid 2.0': Money-Making AI Strategy", https://www.khan.co.kr/article/202503032107005
101) SK Telecom Newsroom, 2025, "SK Telecom Releases Open-Source Korean LLM A.Dot X 4.0 (Built on Qwen 2.5)", https://aimatters.co.kr/news-report/english-news/25688

고 성능으로 평가받는 Qwen 2.5(72B) 모델을 기반으로 개발되었다. SK 텔레콤은 여기에 방대한 자체 한국어 데이터를 지속 사전 학습시키며 자체 토크나이저를 적용하여, 한국어 처리 효율을 GPT-4o 대비 33% 향상[102]시켰다. 이러한 기술력은 B2B 상용 모델뿐만 아니라 일부 경량화 버전을 오픈소스로 공개하여 국내 생태계 활성화에도 기여하고 있다. 이와 별도로, 전략적 파트너인 Anthropic과는 통신 업무에 특화된 'Telco LLM'을 공동 개발하여 도이치텔레콤, 싱텔 등 'Global Telco AI Alliance' 회원사들에게 공급하는 글로벌 표준화 전략을 추진 중이다.

AI 인프라 및 B2B 사업에서도 가시적인 성과를 내고 있다. 글로벌 AI 데이터센터 사업을 통해 NVIDIA GPU 기반의 클라우드 서비스(GPUaaS)를 제공하며, 2025년까지 서울 가산 등에 대규모 GPU 클러스터를 조성해 'AI 고속도로'를 구축할 계획이다. 또한 A.X 기반의 기업용 솔루션 'A. Biz'를 SK 계열사 및 50여 개 외부 기업에 배포하여 업무 자동화를 지원했고, 그 결과 2024년 AI B2B 매출은 전년 대비 32% 성장하였다.

자사 비즈니스 통합(AIX)과 B2C 서비스 부문에서는 5,000만 가입자 데이터를 학습한 AI 기술이 적용되었다. T맵은 에이닷과 통합되어 대화형 길 안내 기능을 제공하고, AI 고객센터(AICC)는 연간 수천만 달러의 비용 절감을 실현하였다. 개인 비서 앱인 'A.dot(에이닷)'은 통화 녹음/요약, 실시간 통역 등의 킬러 기능을 앞세워 2025년 9월 기준 월간 활성 사용자(MAU) 1,000만 명을 돌파[103]하며 국내 최대의 모바일 AI 플랫폼으로 자리 잡았다.

SK Telecom의 2024년 총 매출은 $15.8B(약 17.94조 원) 이며, 이 중 AI 관련 매출(AIX 및 데이터센터 등)은 전년 대비 19% 성장[104]하여 약 $300M(약 4,000억 원)[105] 규모를 달성한 것으로 추산된다. 특히 AIX(AI Transformation) 사업부는 클라우드와 B2B 솔루션의 호조로 32% 의 높은 연간 성장률(YoY)을 기록[106]하며 수익화의 핵심 동력으로 부상하였다.

에이닷(A.dot) 전화의 AI 기능 도입과 고객센터 자동화를 통해 연간 $50M

102) HelloT, 2025, "SK 텔레콤 unveils A.X 4.0 (KMMLU 78.3 vs GPT-4o 72.5)", https://www.hellot.net/news/article.html?no=10288
103) Asia Economy, 2025, "SK 텔레콤 Aidot Surpasses 10 Million Monthly Users", https://cm.asiae.co.kr/en/article/202510210903317160
104) SK Telecom Press Release, 2025, "SK Telecom Announces FY 2024 Results (AI Revenue +19%)", https://www.SKtelecom.com/en/press/press_detail.do?idx=1631
105) TelecomTV, 2024, "SK Telecom Invested $300M+ for AI Growth", https://www.telecomtv.com/content/telcos-and-ai-channel/what-s-up-with-sk-telecom-lumen-ntt-docomo-51001
106) SK Telecom Press Release, 2025, "SK Telecom Announces FY 2024 Results", https://www.SKTelecom.com/en/press/press_detail.do?idx=1631

이상의 운영 비용 절감 효과[107]를 거두었으며, 2025년에는 AI 관련 매출이 $400M 이상으로 확대될 것으로 전망된다.

국내 모바일 AI 시장에서 통신 기반의 강력한 사용자 접점을 활용하여 압도적인 우위를 점하고 있다. AI 개인 비서 앱 '에이닷(A.dot)'은 2025년 9월 기준 월간 활성 사용자(MAU) 1,000만명 을 돌파[108]하며 국내 스마트폰 사용자의 약 20% 가 사용하는 필수 앱으로 자리 잡았다. 이는 통신사 AI 서비스 중 가장 높은 점유율이다. 글로벌 오픈소스 LLM 생태계에서는 Hugging Face를 통해 공개된 모델들의 누적 다운로드가 약 4.67만 회를 기록[109]하며, 국내 기업 중에서는 중위권의 인지도를 확보하고 있다.

3. 네이버

네이버는 자사의 초대규모 AI인 'HyperCLOVA X'를 중심으로 자체 검색, 쇼핑 서비스 및 퍼블릭 클라우드 서비스 고도화를 추진하며 개방형 생태계를 아우르는 정교한 투트랙(Two-Track) 전략을 구사하고 있다. 기업 고객을 대상으로는 단순한 API 제공을 넘어, 고객사의 데이터센터 내에 폐쇄형 AI 인프라를 직접 구축해 주는 '뉴로클라우드(Neurocloud)'와 퍼블릭 클라우드를 병행하는 '하이브리드 AI' 전략을 통해 시장을 장악하고 있다. 이를 바탕으로 네이버의 클라우드 및 AI 관련 매출은 전년 대비 견조한 성장세를 보이며 연간 약 3억 달러(약 4,000억 원) 규모를 달성한 것으로 추산된다.

그리고, 폐쇄형 전략에만 머물지 않고, 오픈소스 생태계로의 확장을 적극 추진하고 있다. 네이버는 경량화된 오픈소스 모델인 'HyperCLOVA X SEED' 시리즈(예: 14B 모델)[110]를 허깅페이스(Hugging Face) 등을 통해 공개하며 생태계 확장에 나서고 있다. 이는 Llama나 Gemma와 같은 글로벌 오픈 모델에 대응하여 한국어 처리 능력이 뛰어난 소형 언어 모델(sLLM) 확산을 목표로 하고 있다. 이를 통해 외부 개발자가 네이버 AI 기술력을 체험하고

107) Twimbit, 2024, "The Race to AI-Native Telco: SK Telecom Leads the Pack (35% Share)", https://content.twimbit.com/insights/the-race-to-ai-native-telco-sk-telecom-leads-the-pack
108) 매일경제, 2025, "SKT A.dot Surpasses 10 Million MAU", https://www.mk.co.kr/en/it/11446994
109) Hugging Face, 2025, "SK Telecom Organization Profile & Download Stats", https://huggingface.co/SK 텔레콤
110) Naver Cloud (Hugging Face), 2025, "HyperCLOVA X SEED Technical Report", https://huggingface.co/Naver-hyperclovax/HyperCLOVAX-SEED-Think-14B

장기적으로 유료 API 및 클라우드 생태계로 유입시키는 확산 전략이다.

B2C 영역에서는 'On-Service AI' 전략을 통해 검색, 쇼핑, 콘텐츠 등 핵심 플랫폼의 수익성을 극대화한다. 생성형 AI 검색 서비스인 'Cue:(큐:)'는 복잡한 질의 의도를 파악하여 쇼핑 및 로컬 정보와 연동된 답변을 제공함으로써, 2024년 말 기준 누적 사용자 3,000만 명을 돌파[111]하고 전체 검색 트래픽의 15%를 점유하는 핵심 서비스로 자리 잡았다.

커머스 부문에서도 15억 개 이상의 상품 데이터를 분석하는 'AiTEMS'[112] 추천 시스템을 고도화하여 구매 전환율을 도입 초기 대비 2배 가까이 끌어올렸으며, AI 추천 영역의 거래액 비중을 4배 이상 증가시키는 등 실질적인 매출 증대 효과를 입증하였다. 또한, 글로벌 시장에서 네이버 웹툰의 AI 번역 기술이 창작 및 유통 속도를 10배 이상 가속화하고 있으며, 라인(Line)의 AI 어시스턴트 기능과 결합하여 아시아 시장 내 AI 영향력을 확대[113]하고 있다.

네이버의 2024년 연결 기준 총매출은 전년 대비 11.0% 성장한 10조 7,377억 원(약 75억 달러)을 기록[114]하며 사상 최대 실적을 달성하였다. 특히 연간 영업이익은 1조 9,793억 원으로 전년 대비 32.9% 급증했는데, 이는 AI 도입에 따른 운영 효율화와 광고 타겟팅 고도화가 수익성 개선에 크게 기여한 결과다.

이 중 AI 관련 매출(AIX 및 클라우드 포함)은 약 $300M(약 4,200억 원) 규모로 추산되며, Cue: AI 검색과 HyperCLOVA X의 B2B 계약 확대가 주요 성장 동력으로 작용하고 있다. 2025년에는 AI B2B 매출이 본격화되면서 해당 분야 매출이 $600M 수준으로 두 배 가까이 성장할 것으로 전망[115]된다.

국내 검색 시장에서 구글의 추격 속에서도 2025년 기준 약 48%~50%의 점유율을 방어하며 1위를 유지[116]하고 있다. 이는 생성형 AI 검색 'Cue:'의 도

111) Maeil Business Newspaper, 2025, "Naver AI Search Users Exceed 30 Million", https://www.mk.co.kr/en/it/11460857
112) Pickool, 2025, "Naver AI Shopping Platform Doubles Conversion Rates", https://www.thepickool.com/네이버-ai-shopping-platform-doubles-conversion-rates/
113) ElectroIQ, 2025, "Naver Statistics By Revenue and Facts (2025)", https://electroiq.com/stats/Naver-statistics
114) Naver Corp Press Release, 2025, "Naver Achieves Record Annual Revenue of 10.7 Trillion Won in 2024", https://www.Navercorp.com/media/pressReleasesDetail?seq=32292
115) 연합뉴스, 2025, "Naver forecast to post record annual revenue, AI Profitability Improving", https://en.yna.co.kr/view/AEN20250120006300320
116) Matrix BCG, 2025, "Competitive Landscape of Naver: Market Share & Financials", https://matrixbcg.com/blogs/competitors/Naver

입이 사용자 체류 시간을 늘리고 검색 만족도를 높인 덕분으로 분석되고 있다. 글로벌 시장에서 자회사 라인 야후를 통해 일본 내 메신저 시장 점유율 78% 이상을 차지[117]하고 있으며, 월간 활성 사용자(MAU) 9,700만 명을 기반으로 AI 챗봇 및 비즈니스 솔루션 시장에서 독보적인 경쟁력을 발휘하고 있다.

미래 경쟁력 확보를 위한 R&D 투자는 매출의 20~25% 수준인 연간 약 2조 원(약 $1.4B) 규모[118]에 육박한다. 이는 주로 초거대 AI 모델 고도화, AI 경량화 기술, 그리고 각(GAK) 세종 데이터센터를 포함한 AI 인프라 확충에 집중되고 있다. 특히 2025년에는 소버린 AI 생태계 주도를 위해 국내 AI 생태계 육성에 향후 6년간 1조 원 규모의 펀드를 조성하는 등 인프라 및 생태계 확장에 대한 투자를 더욱 가속화할 계획이다.

4. 업스테이지

업스테이지는 자사의 핵심 LLM인 'Solar'를 Apache 2.0 라이선스로 공개[119]하여, 외부 기업들이 별도의 제약 없이 상업적으로 활용할 수 있도록 하는 개방형 생태계 전략을 취하고 있다. 특히 주력 모델인 Solar 10.7B는 독자적인 DUS(Depth Up-Scaling) 기술을 적용하여 100억 개 수준의 파라미터로 300억 개 이상의 대형 모델(Mixtral 8x7B 등)을 능가하는 성능 효율성을 달성[120]하였다. 이러한 기술력으로 허깅페이스 개방형(Open) LLM 리더보드에서 1위를 차지하였다.

주요 상용화 성과로는 Document AI 솔루션으로 삼성생명은 이 솔루션을 도입하여 진료비 영수증 등 복잡한 보험 청구 서류 7종에 대해 95% 이상의 인식 정확도를 달성했으며, 이를 통해 수작업 대비 비용과 시간을 82%까지 절감하는 효과를 거두었다. 또한 신한은행, 한국투자증권 등 금융권과 LG유플러스 등 통신사를 포함하여 2024년 기준 300개 이상의 기업 고객을 확보하며 B2B 시장에서 실질적인 매출을 창출[121]하고

117) TAMLO, 2025, "LINE User Trends 2025: The Largest Messaging App in Japan (97M MAU)", https://tam-tamlo.com/en/307
118) Chosun Biz, 2025, "Naver invests over 1 trillion won in R&D in H1, Focusing on AI", https://biz.chosun.com/en/en-it/2025/08/19/PBOO5KCZ3RAY7E7NY2WHE5I57Q/
119) Hugging Face, 2025, "Upstage SOLAR 10.7B Apache 2.0 License & Performance", https://huggingface.co/Upstage/SOLAR-10.7B-Instruct-v1.0
120) Upstage Press Release, 2025, "Solar 10.7B Tops Global LLM Leaderboard, Beating Mixtral", https://Upstage.ai/news/solar-10-7b-emerges-as-worlds-top-pre-trained-llm
121) Samsung SDS Case Study, 2025, "Samsung Life Insurance Achieves 95% OCR Accuracy with

있다. Document AI는 단순 텍스트 인식을 넘어 표, 체크박스, 손글씨까지 99% 수준으로 정교하게 인식하여 금융 및 법률 분야의 업무 자동화(RPA)에 핵심적인 역할을 수행하고 있다.

신사업 영역에서 'Solar API' 플랫폼을 통해 개발자 생태계를 확장하고 있으며, 미국 법인 설립과 함께 아마존 웹 서비스(AWS) 마켓플레이스에 입점하여 글로벌 판매[122] 채널을 다각화하고 있다. 2025년에 일본 및 동남아시아 시장 진출을 본격화하면서 프라이빗 LLM 구축 및 AI 에이전트 개발 도구 제공에 주력하며 수익 모델을 고도화하고 있다.

업스테이지의 2024년 총매출은 전년 대비 150% 성장한 약 $21M(약 280억 원) 수준으로 추정되며, 이는 주력 제품인 'Document AI'의 공급 확대와 'Solar API'의 사용량 증가에 힘입은 결과다. 2025년에는 글로벌 B2B 매출이 본격화됨에 따라 매출이 $40M(약 530억 원) 이상으로 두 배 가까이 증가할 것으로 전망된다. 재무적 안정성을 확보하기 위한 투자 유치 활동도 활발히 진행되어,

2025년 8월 Series B Bridge 펀딩을 통해 $45M(약 600억 원)을 추가로 확보[123]하였다. 이로써 누적 투자금은 $157M(약 2,100억 원)에 달하며, 기업 가치는 기존 4,000억 원대에서 약 2배 상승한 ₩790B(약 $590M) 규모로 평가[124]받고 있다. 이번 라운드에는 KDB산업은행, Amazon Web Services (AWS), AMD 등 글로벌 빅테크 기업들이 전략적 투자자(SI)로 새롭게 합류하여 글로벌 시장 확장에 대한 기대감을 높였다.

국내 문서 AI 시장에서 약 40%의 점유율로 1위를 하고 있으며, 특히 데이터 보안과 정확도가 필수적인 보험 및 금융권에서는 60% 이상의 독보적인 점유율[125]을 차지하고 있다. 또한 글로벌 오픈소스 생태계에서 영향력을 확보하고 있다. Solar LLM 시리즈는 Hugging Face 리더보드에서 1위를 차지하며 누적 다운로드 50만 회를 돌파하며 한국 기업 중 가장 높은 글로벌 인지도를 확보하였다.

이러한 기술적 성과와 시장 점유율을 바탕으로 업스테이지는 코스닥 상장을 위한 준비 단계(Pre-IPO)[126]에 진입하였으며, AWS와의 협력을 통해 아시아태평양 및 미국 시장

Upstage Document AI", https://Upstage.ai/news/documentai-samsunglife

122) SiliconAngle, 2025, "Upstage Raises $72M for Global Expansion of Document AI (99% Accuracy)", https://siliconangle.com/2024/04/16/document-ai-startup-업스테이지-raises-72m-global-expansion/

123) Upstage Press Release, 2025, "Upstage Raises $45M in Series B Bridge, Backed by Amazon & AMD", https://www.superbcrew.com/Upstage-raises-45-million-in-series-b-bridge-funding-round/

124) Bloomberg, 2025, "Amazon Joins AMD to Back South Korean AI Startup Upstage", https://www.bloomberg.com/news/articles/2025-08-20/amazon-joins-amd-to-back-south-korean-ai-startup-Upstage

125) CB Insights & Upstage Blog, 2024, "업Upstage named to CB Insights AI 100 2025 (Document AI Share 60%+)", https://Upstage.ai/blog/en/Upstage-named-to-cb-insights-ai-100-2025

진출을 추진하고 있다.

5. 엔씨AI

국내 주요 게임 기업 중 하나인 엔씨 소프트는 자사의 핵심 LLM인 VARCO(바르코)를 현재 기술 공유 및 연구 생태계 기여 목적으로 비상업적 라이선스(CC BY-NC 4.0) 형태로 공개하고 있다.[127] 2025년 2월에 AI 전문 자회사인 엔씨AI를 설립[128]하여 본격적인 상업화에 나서고 있으며. 이 신규 법인은 한국, 일본, 동남아시아의 중소 게임 개발사를 주요 타겟으로 삼아, Unity나 Unreal Engine과 같은 게임 엔진에서 즉시 활용 가능한 '플러그인 형태의 AI 개발 도구'를 B2B 솔루션으로 제공할 계획이다.

이미 사내 적용을 통해 검증된 기술을 기반으로 VARCO Studio라는 상업 솔루션을 제공하고 있다. 이는 시나리오, 캐릭터 원화, 퀘스트 등을 생성하는 올인원 저작 도구로, 실제 사내 게임 개발 파이프라인에 도입되어 개발 기간을 평균 30% 단축시키고 연간 약 2,000만 달러(약 260억 원) 이상의 비용 절감 효과[129]를 거두고 있다.

또한, 게임 용어에 특화된 번역 모델인 'VARCO-MT'는 기존 인력 중심 번역 대비 비용을 80% 절감[130]하면서도 캐릭터의 말투와 맥락을 살린 고품질 번역을 제공하여 글로벌 시장에서 세부 시장별 지역화 효율성을 극대화하고 있다. 더불어, AI 에이전트가 실제 유저처럼 게임을 플레이하며 버그를 찾아내는 자동화된 QA(품질 보증) 솔루션도 개발[131]되어, 테스트 비용 절감과 출시 전 안정성 확보를 추진하고 있다.

자사 게임 통합 사례로는 리니지W의 'NPC 대화 자동 생성' 및 '동적 퀘스트 시스템', 블레이드&소울 2의 'AI 기반 캐릭터 행동 패턴', 그리고 최신작 Throne and Liberty(TL)의 'LLM 기반 대화 시스템' 등이 있으며, 2024년 기준 10여 개의 타이틀에 VARCO 기술[132]이 적용되어 플레이어들에게 혁신적인 경험을 제공하고 있다.

---

126) Growjo, 2025, "Upstage Revenue Estimate & Funding Data", https://growjo.com/company/Upstage
127) NC Research, 2025, "VARCO LLM License (CC BY-NC 4.0) & Technical Report", https://ncsoft.github.io/ncresearch/varco/
128) NCSOFT News, 2024, "Establishment of 'NC AI' Spinoff for Commercialization", https://about.ncsoft.com/en/news/article/news_update_241128
129) NCSOFT IR, 2025, "Q2 2025 Earnings Release: Operating Profit Turnaround", https://about.ncsoft.com/en/news/article/news_update_250812
130) Reddit & Industry Reports, 2024, "Game Localization Cost Reduction with AI (80% Saving)", https://www.reddit.com/r/gamedev/comments/636poq/save_money_on_game_localization_1_real_prices
131) Tech Wave Arena, 2024, "AI QA Automation: Reducing Costs and Time", https://techwavearena.com/automatically-find-bugs-on-your-website-save-time-and-reduce-costs-with-fix-ai-no-code-qa-solution/
132) 매일경제 2024, "NCSoft Commercializes VARCO for External Content Companies in 2025",

엔씨AI의 자체 매출은 아직 집계되어 있지 않으나, 엔씨소프트의 2024년 총매출[133]은 전년 대비 감소한 약 1조 5,800억 원(약 $1.15B)을 기록했으며, 일회성 비용과 마케팅 투자 증가로 영업손실을 기록하였다. 이러한 상황에서 엔씨AI의 AI 기술은 사내 개발 파이프라인에 적용하여 내부 생산성과 효율성을 향상시키며 연간 $20M(약 260억 원) 이상의 비용 절감 효과를 얻었을 것으로 추정된다.

특히 'VARCO Studio' 도입을 통한 개발 기간 30% 단축 효과는 차기작 'Aion 2' 등의 출시 속도를 높이는 핵심 동력이 되고 있으며, AI 기반 NPC 대화 시스템은 플레이어의 몰입도를 높여 만족도를 약 25% 향상[134]시킨 것으로 분석된다. 국내 게임 AI 분야에서 약 70%의 높은 시장 점유율을 유지하며 기술적 우위를 점하고 있다.

넷마블, 넥슨 등 경쟁사들도 AI 도입에 적극적이지만, 자체 LLM을 보유하고 상용화 단계까지 도달한 곳은 엔씨가 유일하다. 글로벌 오픈소스 생태계에서 허깅페이스를 통해 공개한 'Llama-VARCO' 모델 등이 누적 다운로드 약 2,000회~3,000회 수준을 기록[135]하고 있다. 이는 수십만 회를 기록하는 범용 모델들에 비해 적은 수치이나, '게임 특화(Niche Market)' 전략에 집중한 결과로 해석된다. NC Soft는 이러한 특화 데이터를 바탕으로 2030년까지 $51B(약 70조 원) 규모[136]로 성장할 미디어 AI 시장을 정조준하고 있다.

---

https://pulse.mk.co.kr/news/english/11186398

133) NCSOFT IR, 2025, "Annual Financial Results 2024: Revenue KRW 1.58T", https://about.ncsoft.com/en/news/article/news-update-250212

134) Asiae, 2025, "NC AI Evolves to Vertical AI: AI NPC & Multilingual Chat", https://cm.asiae.co.kr/en/article/2025062416210472051

135) Hugging Face, 2025, "NCSOFT Model Download Stats", https://huggingface.co/NCSOFT

136) 중앙데일리, 2025, "NC AI Vision: Boosting K-Content with VARCO (Media AI Market $51B)", https://koreajoongangdaily.joins.com/news/2025-10-08/business/industry/NC-AI-unveils-vision-to-boost-Kcontent-with-gamingdriven-AI/20251007

〈표 5〉 국내 주요 오픈소스AI 기업 현황 요약

구분	주요 오픈소스 AI	법인 유형	AI 관련 주요 사업	오픈소스 전략 목적	주요 파트너십
LG AI	EXAONE 3.0 (7.8B) EXAONE Universe	대기업 연구 조직	제조 산업 특화 (가전, 화학, 신약 등) ChatEXAONE (사내용)	기술력 입증	LG 화학/에너지 솔루션
SK 텔레콤	A.X (에이닷 엑스)	대기업	스마트폰 AI 서비스 (AI 비서, 통화 번역) 고객서비스 향상 및 효율화	글로벌 통신사 LLM 표준화 추구	OpenAI, Anthropic, 도이치 텔레콤, 싱텔
네이버	HyperCLOVA X HyperCLOVA X SEED	대기업	검색(Cue:) 강화 및 쇼핑 추천, 클라우드 서비스	전문가 생태계 활성화, 기술력 입증	인텔 (AI 반도체), 사우디 아람코
업스테이지	Solar Pro (10.7B/22B) Solar Mini	스타트업	문서 특화, AI 서비스 및 사설 LLM 구축	기술력 입증 -> 글로벌 인지도 향상	AWS, 삼성생명, 신한은행
엔씨AI	VARCO LLM VARCO Art/Text	대기업 자회사	게임 특화, AI NPC, QA 자동화	기술력 입증	Unity/Unreal 엔진 (3D 플러그인)

(자체 작성)

제4절 글로벌 주요 오픈소스AI 기술(오픈소스 모델) 동향

2025년 현재, 오픈소스AI 생태계는 과거 BERT(2018)와 GPT-2(2019)로 대표되던 단순 텍스트 생성 단계를 넘어, 산업 전반을 지탱하는 “플랫폼 기술”로 완전히 진화하고 있다. 전통적인 글로벌 빅테크 기업(메타, Google)이 주도하던 생태계에서 중국 기업인 알리바바와 딥시크가 부상하고 있으며, 국내 기업들도 초기 단계이지만 글로벌 경쟁력 확보를 위해 노력하고 있다.

특히 단순 기술 발전에 머무르지 않고 오픈소스 기반 기술 경쟁력을 확보하고 이를 통해 수익화 단계를 추진하면서 산업적 영향력을 확대하고 있다. 특히 기술적으로 AI 기술의 효율성 제고를 통해 기존 상용AI 서비스의 대안으로 주목받고 있다.

초대형 모델의 경량화에 있어서 Llama 4, Qwen 2.5, DeepSeek-V3 등 70B~2T급 모델들은 MoE(Mixture-of-Experts) 및 MLA(Multi-head Latent Attention) 아키텍처를 표준으로 채택하였다. 이를 통해 추론 시 활성화되는 파라미터를 최소화하고 동급 성능의 기존 모델 대비 추론 비용을 획기적으로 절감하고 있다.

특화 연산(Reasoning) 강화 측면에 있어서도 범용 학습을 넘어 CodeLlama, DeepSeek-R1, Solar Pro 등 코드・수학・논리 추론(Chain-of-Thought)에 특화된 모델 구조가 정착되고 있으며 SWE-Bench, GSM8K 등 고난이도 벤치마크 성능이 비약적으로 향상되고 있다[137].

또한 단순 기술에 머무르지 않고 서비스 융합 및 에이전트화를 통해 수익화 단계로 넘어서고 있다. AI 모델이 단순한 질의응답 도구를 넘어, 실제 산업 현장에서 도구를 사용하고 복합적인 업무를 수행하는 에이전트(Agent) 형태로 진화되고 있다. 특히 멀티모달 LLM을 통해 텍스트, 이미지(Vision), 문서(OCR), 영상(Video) 등의 다양한 기능들이 제공되면서 하나의 모델로 복합 기능을 처리하는 멀티모달 네이티브 구조가 보편화되고 AI 서비스 제공 및 타 기능(웹 검색) 연동 등의 에이전트형 서비스가 가능하게 되었다.

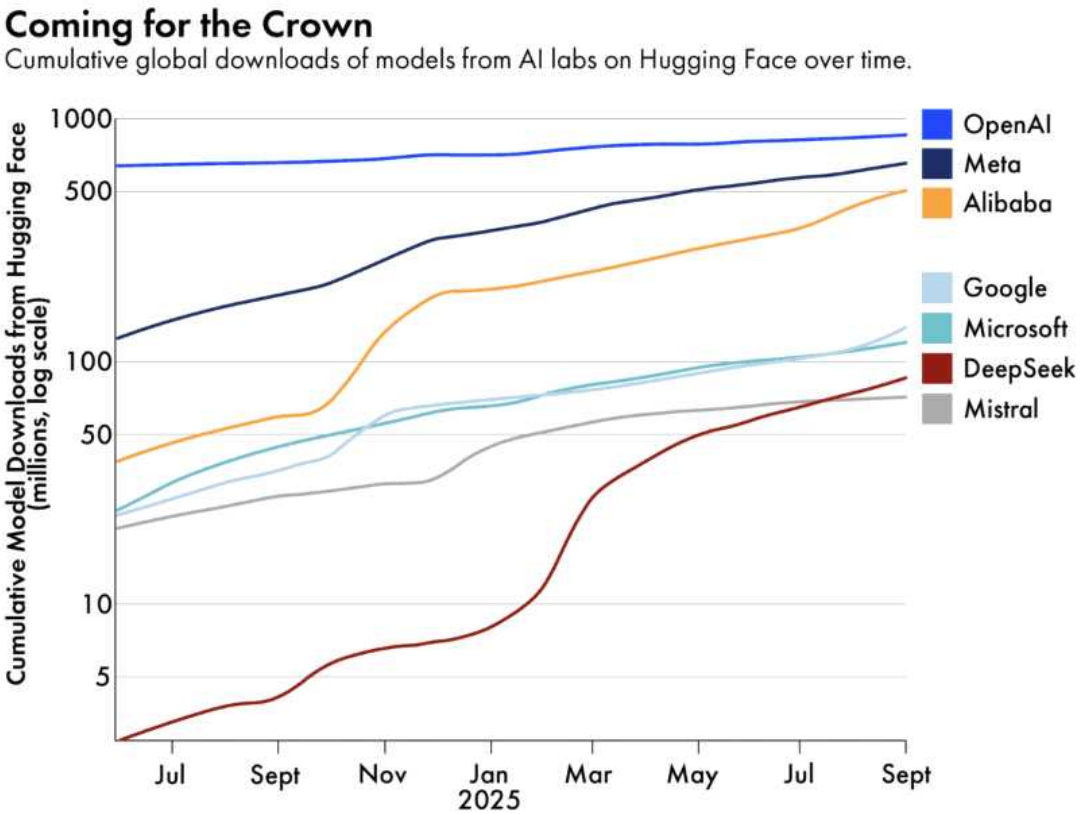
그리고, 최적화 기술을 적용하여 거대 파운데이션 모델에서 벗어나 sLLM 경량 모델(젬마3, 큐웬 1.5B, Mistral 7B)과 QLoRA, Adapter 등 효율적 튜닝 기술과 결합되어 온디바이스 및 에지 최적화를 통해 스마트카, 로봇, IoT 기기 등 저사양 엣지(Edge) 환경으로 빠르게 확산되고 있다.

---

137) LMSYS Org, 2025, Chatbot Arena Leaderboard: Coding & Hard Prompts, https://chat.lmsys.org

이렇게 오픈소스AI 기술들이 전세계적으로 주목받으며 다양한 산업 분야로 확산되고 있다. 특히 글로벌 오픈소스 생태계가 소수 빅테크 주도의 고비용 독점에서 누구나 접근 가능한 저비용 보편화로 전환되는데 큰 기여를 하면서 보편적 AI 시대로의 전환에 오픈소스AI 기술들이 크게 기여하고 있다.

[그림 21] Coming for the Crown“: 글로벌 AI 모델 생태계의 지각변동 현황



(출처) (The Wire China) Cheap and Open Source, Chinese AI Models Are Taking Off 2025.11.

따라서, 앞에서 살펴본 국내외 오픈소스AI 기업들의 주요 모델들에 대한 조사를 통해 글로벌 오픈소스AI 기술 현황에 대해 파악해보고자 한다.

1. 메타 라마 (Llama)

라마(Llama)는 메타(메타, 구 Facebook)가 개발한 오픈 웨이트 대규모 언어모델(LLM) 시리즈로, 2023년 2월 첫 버전 공개 이후[138] 전 세계 오픈소스 AI 생태계

의 사실상 표준 모델로 자리잡았다. 라마는 단순히 모델 가중치만 공개하는 오픈 웨이트 접근법을 통해 연구자와 기업이 자유롭게 커스터마이징하고 파인튜닝할 수 있는 환경을 제공하여 현재까지 누적 10억 회 이상 다운로드[139] 되는 등 글로벌 AI 커뮤니티에서 가장 널리 확산된 LLM이다.

메타는 라마를 통해 폐쇄형 모델(GPT, Claude 등) 중심의 AI 시장에서 오픈소스 진영의 대항마로 자리매김하며, 투명성・접근성・비용 효율성을 강조한 AI 민주화 전략을 추진해오고 있다. 2025년 중국의 DeepSeek・Qwen 등이 Llama 3.1 405B에 필적하거나 능가하는 성능을 훨씬 낮은 비용과 더 관대한 라이선스(Apache 2.0, MIT)로 제공하면서, 오픈소스 AI 시장의 지각변동은 라마의 위상에 중대한 도전을 가하고 있다.

ATOM Project 보고서(2025년 8월)에 따르면, Hugging Face에서 중국 모델의 월별 파생 모델 비중이 40% 이상을 차지하는 반면 라마의 점유율은 2024년 가을 50%에서 2025년 15%로 급락[140]하였다. LMArena 상위 10개 오픈 모델 전체가 중국 기관 개발 모델로 채워졌으며, ArtificialAnalysis 순위에서도 상위 3개가 모두 중국산 모델이었다.

이러한 위기 속에서 메타는 2025년 4월 5일 Llama 4를 긴급 공개[141]하였다. 업계 전문가들은 DeepSeek R1・V3의 성공이 메타를 “패닉 버튼“을 누르게 만들었으며, 메타가 “전쟁실(war room / TF팀)” 을 소집해 DeepSeek의 저비용 고성능 전략을 분석하였다고 보도[142]하였다. Llama 4는 Scout・Maverick・Behemoth 세 가지 모델로 구성되며, 최초로 MoE(Mixture-of-Experts) 아키텍처와 네이티브 멀티모달 기능을 도입해 “새로운 시대의 시작“을 발표하였다.

하지만 Llama 4는 OSI(Open Source Initiative) 기준 10개 조건 중 9개를 충족하지 못하며, “오픈소스“가 아닌 통제된 “오픈 웨이트“ 모델이라는 비판에 직면하

138) Meta AI Blog, 2024, With 10x growth since 2023, Llama is the leading engine of AI innovation, https://ai.meta.com/blog/llama-usage-doubled-may-through-july-2024/
139) OpenTools, 2025, Meta's Llama AI Hits 1.2 Billion Downloads: A New Milestone in Open-Source AI, https://opentools.ai/news/메타s-llama-ai-hits-12-billion-downloads-a-new-milestone-in-open-source-ai
140) Business Standard, 2025, How much of Silicon Valley's AI boom is powered by China's models, https://www.business-standard.com/world-news/how-much-of-silicon-valley-s-ai-boom-is-powered-by-china-s-models-125111000156_1.html
141) Interconnects.ai, 2025, Llama 4: Did Meta just push the panic button?, https://www.interconnects.ai/p/llama-4
142) Gavin Simpson (LinkedIn Pulse), 2025, Llama-3.1 (405B) vs. DeepSeek-v3 (671B) on 8x MI300X, https://www.linkedin.com/posts/gavinssimpson_llama-31-405b-vs-deepseek-v3-671b-on-activity-7318319056518844416-LN_o

였다. EU에서는 규제 회피를 위해 접근이 전면 차단되었으며, 상업적 이용 시 MAU 7억 명 제한·지리적 제약·재배포 조건 등이 포함된 Llama Community License로 인해 진정한 오픈소스와 거리가 있다는 지적이 제기[143]되었다.

또한 메타는 Llama 4를 기반으로 한 유료 API 서비스를 도입[144]하며, Groq·Fireworks AI 등 파트너를 통해 Scout $0.11~0.34/1M 토큰, Maverick $0.27~0.85/1M 토큰의 가격으로 제공하기 시작하였다. 이는 기존 "완전 무료 오픈소스" 이미지와 충돌하며 "클로즈드 전략"으로의 전환 가능성을 시사했고, 일부 전문가는 메타가 "오픈소스 레이블을 납치하였다"며 "오픈워싱(open-washing)"이라 비난하였다.

이런 논란 속에서도 라마는 다양한 파라미터 규모(1B~2T급)의 모델 패밀리를 제공해 온디바이스부터 클라우드 엔터프라이즈급 워크로드까지 모두 커버하며, AWS·Azure·Google Cloud 등 주요 클라우드 사업자와 긴밀히 협력해 엔터프라이즈 채택을 가속화하고 있다.

<표 6> 라마 주요 모델

버전	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	특징 및 혁신
Llama 1	7B/13B/33B/65B	LLM 오픈소스화의 시발점 - 연구 목적 제한 배포	2K	연구 목적, LLM 오픈소스화
Llama 2	7B/13B/70B	- 상업적 이용 허용 (Community License) - RLHF 대규모 적용 및 안전성 강화	2K	상업적 허용, RLHF 대규모 적용
Llama 3	8B/70B	- 대형화 및 인스트럭션 튜닝 고도화 - 15조 토큰 학습 데이터	128K	대형화, 인스트럭션 튜닝, 15조 토큰
Llama 3.1	8B/70B/405B	405B 초대형 모델 공개	128K	멀티링구얄, FP8,

143) LlamaIMoel.com, 2025, Llama AI for Commercial Use: License & Enterprise Impact, https://llamaimodel.com/commercial-use/

144) Groq, 2025, Llama 4 Live Today on Groq — Build Fast at the Lowest Cost, https://groq.humain.ai/llama-4-now-live-on-groq-build-fast-at-the-lowest-cost-without-compromise/

		(오픈 웨이트) - 다국어 처리 강화 및 FP8 양자화		405B 오픈웨이트
Llama 3.2	1B/3B/11B/90B	- 경량화(SLM) 및 멀티모달(Vision) 통합 - 모바일/엣지 최적화	128K	경량화・멀티모달・모바일・비전
Llama 3.3	70B		128K	포스트 트레이닝, GPT-4급 효율화
Llama 4	Scout(17B MoE), Maverick(400B MoE), Behemoth(~2T)	- MoE(Mixture-of-Experts) 아키텍처 도입 - Native Multimodal (텍스트・이미지・비디오) - 17B Active Params (Scout)	1M~10M	Sparse MoE, 장문 멀티모달(텍・이미지・비디오), 200개 언어

(자체 작성)

메타의 Llama(Large Language Model Meta AI) 시리즈[145]는 2023년 2월 첫 버전(Llama 1) 공개 이후 전 세계 오픈소스 AI 생태계의 '사실상 표준(De Facto Standard)'으로 자리 잡았다. 기본적으로 텍스트 생성 중심의 대규모 언어모델(LLM)에서 출발했으나, 최근 버전에서는 멀티모달・코드・추론 등 다양한 특화 기능을 통합하며 범용 플랫폼으로 진화하였다.

특히 Llama 3 이후의 모든 버전은 사전 학습(Pre-trained) 모델과 지시 이행(Instruction Tuned) 모델을 동시에 제공한다. 이를 통해 연구자는 기초 모델을 자유롭게 커스터마이징하고, 기업은 즉시 사용 가능한 챗봇 및 에이전트 구축에 활용할 수 있는 환경이 조성되었다.

1) 초기 기반 구축 및 생태계 확장 (2023년)

Llama 1 (2023.02)는 7B・13B・33B・65B의 4가지 파라미터 규모로 공개되었다. 당시에는 연구 목적(Non-commercial)으로만 제한 배포되었으나, 고성능 LLM의 민주화 가능성을 입증한 효시가 되었다.

뒤이어, Llama 2 (2023.07)는 7B・13B・70B 세 가지 규모로 출시되었으며, 상업

145) Wikipedia, 2025, Llama (language model), https://en.wikipedia.org/wiki/Llama_(language_model) (Llama 시리즈의 전체 역사 및 버전별 스펙 개요)

적 이용을 허용하는 'Llama 2 Community License가 도입되면서 기업들의 본격적인 채택이 시작되었다.

2) 성능 고도화 및 엔터프라이즈 최적화 (2024년)[146]

Llama 3 (2024.04)는 8B・70B 두 가지 규모로 공개되었다. 약 15조 토큰의 대규모 학습 데이터와 1,000만 건 이상의 인간 피드백(RLHF)을 반영하여 지시 이행 성능을 대폭 향상시켰다. 특히 Llama 3.1 (2024.07)는 : 시리즈 최초로 405B 초대형 모델을 포함해 8B・70B・405B로 확장되었으며, 당시 세계 최대 규모의 오픈 웨이트 모델로 평가받았다. 128K 토큰의 컨텍스트 윈도우, 다국어 처리 강화, 8비트 양자화(FP8) 지원 등 엔터프라이즈 워크로드에 최적화된 기능들이 대거 탑재되었다. 특히 코드 생성・수학 추론・멀티턴 대화에서 GPT-4와 대등하게 경쟁할 수 있는 수준에 도달하였다.

Llama 3.2 (2024.09)는 온디바이스(On-device) 시장을 겨냥한 1B・3B 경량 모델과, 비전-언어 통합 기능을 갖춘 11B・90B 멀티모달 모델을 추가하여 응용 범위를 모바일과 엣지 디바이스로 확장[147]하였다. Llama 3.3 (2024.12)는 포스트 트레이닝 기술을 대폭 개선한 70B 단일 모델이다. 기존 3.1 버전의 405B 모델에 필적하는 고성능을 더 작은 모델 사이즈로 구현하여, 실질적인 운영 비용 효율성을 극대화하였다.

3) 아키텍처 혁신과 차세대 도약 (2025년)[148]

Llama 4 (2025.04.05.)는 DeepSeek 등 경쟁 모델의 부상에 대응하여 출시된 차세대 모델군이다. 시리즈 최초로 MoE(Mixture-of-Experts) 아키텍처와 네이티브 멀티모달(텍스트・이미지・비디오) 기능을 통합하며 기술적 도약[149]을 이뤘다. 세부 보델은 3가지로 제공되며 Scou 모델은 활성 파라미터 17B 수준의 고효율 모델로, 업계 최대 수준인 1,000만 토큰의 컨텍스트를 지원한다. Maverick은 400B 파라미터 규모의 고성능 모델로, 100만 토큰 컨텍스트와 200개 언어를 지원한다.

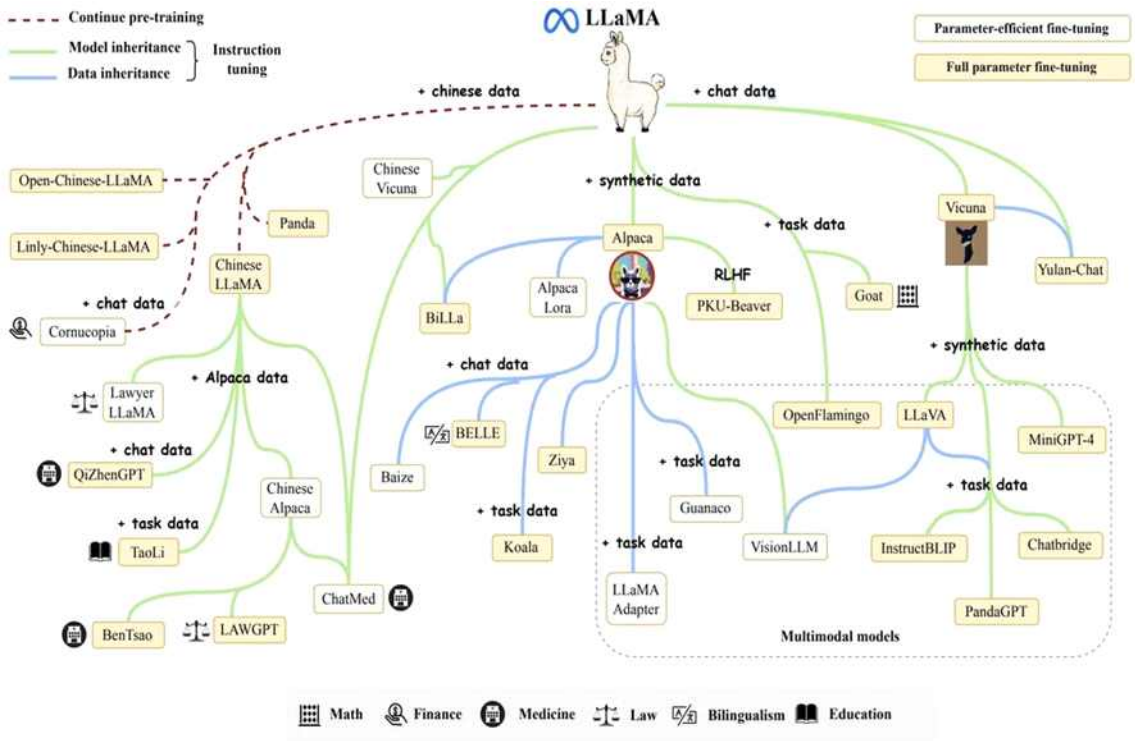
146) Meta AI, 2024, Introducing 메타 Llama 3: The most capable openly available LLM to date, https://ai.meata.com/blog/Meta-llama-3/
147) Moor Insights & Strategy, 2025, Meta Llama's Enterprise AI Value, https://moorinsightsstrategy.com/메타-llama-enterprise-value/
148) Interconnects.ai, 2025, Llama 4: Did Meta push panic button?, https://interconnects.ai/p/llama-4-panic
149) TechCrunch, 2025, 메타 releases Llama 4, https://techcrunch.com/2025/04/04/메타-llama-4-release

Behemoth은 AGI급 성능을 목표로 학습 중인 초대형 모델로 예고되었다.

2.3 파생 모델 생태계[150]

라마는 Hugging Face 기준 20만 개 이상의 파생 모델(전체 LLM 파생의 35%), 일평균 100만 회 이상 다운로드를 기록하며 다양한 도메인에 맞춤형 인스트럭션 튜닝, 전문 챗봇, 코드 생성, 다국어, 모바일·엣지 실행 등 혁신적 활용을 이루었다.

[그림 22] 라마 파생 모델 생태계



(출처 : https://arxiv.org/abs/2504.127370 Chinese-Vicuna: A Chinese Instruction-following Llama-based Model

<표 7> 메타 라마 개요

개발자	메타	홈페이지	https://www.llama.com/
최초공개 시기	2023.02.	라이선스	LLAMA 3.3
주요 특징	오픈웨이트 대표 / 8B~405B~400B(MoE) / 멀티모달 / 엔터프라이즈 표준 / 산업/학술 확장		

150) Maginative, 2025, Meta's Llama Hits 1B Downloads, https://maginative.com/article/메타s-llama-1-billion-downloads/

<table>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/메타-llama</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>591</td><td>15</td><td>70</td><td colspan="2">11</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Llama-4-Maverick (400B MoE)</td><td>61,442</td><td>2,508</td><td>60,031</td><td>139</td></tr>
<tr><td rowspan="5">주요 활용 사례</td><td colspan="2">(PWC) 감사・세무・규제 문서 분석, 규제 준수 자동화</td><td colspan="3">https://www.llama.com/resources/case-studies/pw</td></tr>
<tr><td colspan="2">(Sofya)임상 어시스턴트, 진단・치료 계획 자동 생성</td><td colspan="3">https://www.llama.com/resources/case-studies/sofya</td></tr>
<tr><td colspan="2">(Brain4Data)규제 산업 신뢰성/출처 추적 솔루션</td><td colspan="3">https://www.llama.com/resources/case-studies/brain4data</td></tr>
<tr><td colspan="2">(Shopify)상품 메타데이터 자동 생성 및 실제 대규모 AI 추론</td><td colspan="3">https://www.llama.com/resources/case-studies/shopify</td></tr>
<tr><td colspan="2">(CodeGPT)AI 코딩 어시스턴트, 개발 효율화</td><td colspan="3">https://www.llama.com/resources/case-studies/codegpt</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스
- 모델 가중치는 Open Weight(일부 Behemoth 미공개), HuggingFace・메타 공식・클라우드 다운로드로 공개 코드는 추론/배포 코드만 공개되며, 사전학습 파이프라인/데이터는 비공개

- 데이터셋은 학습 데이터 일부만 “publicly available sources“로 명시

- 공개 라이선스는 Llama Community License(상업적 MAU 7억 제한, Acceptable Use Policy)이며, OSI 기준 미부합(LinkedIn, 2025년 6월), EU 차단이 적용</td></tr>
</table>

(자체 작성)

2. 알리바바 Qwen

알리바바 클라우드(Alibaba Cloud)가 개발한 Qwen 시리즈는 2023년 첫 공개 이후, 2025년 현재(11월 기준) 글로벌 누적 다운로드 6억 건 이상, 파생 모델 17만 개 이상을 기록하며 명실상부한 “아시아 최상위 오픈소스 모델”로 자리 잡았다. Qwen은 119개 언어를 지원하는 다국어 능력과 기업 및 개인에게 무료로 제공되는 강력한 오픈소스 정책을 기반으로 GPT-4o, Claude 3.5, DeepSeek 등 서구권 및 자국 경쟁 모델들과 어깨를 나란히 하고 있다.

특히 최신 Qwen 2.5-Max와 Qwen 3 시리즈[151]는 약 20~23조 토큰의 초대규모 데이터 학습과 MoE(Mixture-of-Experts) 아키텍처를 도입하여, 코딩・수학・추론・장문 이해 등 모든 영역에서 오픈소스 최고 수준의 성능을 구현[152]하였다. 또한, 2025년 11월 출시된 공식 AI 에이전트 ‘Qwen 앱은 출시 첫 주 만에 1,000만 다운로드를 돌파[153]하며 클라우드와 온디바이스를 아우르는 “차세대 AI 슈퍼앱”으로 부상하고 있다.

<표 8> 큐웬 주요 모델

모델명	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도
Qwen 1.0 (2023.04)	7B/14B	- Transformer 기반의 대규모 사전 학습 - 중국어/영어 이중 언어 최적화	8K	중국 최초 대규모 오픈소스 LLM
Qwen 1.5 (2023.07)	2.4B/7B/14B	- 경량화(Small) 모델 라인업 추가 - 초기 코드(Code) 및 멀티모달 기능 실험	8K	경량화/멀티모달/코드 초기
Qwen 2.0-7B (2024.01)	7B	- 추론 속도 및 모바일 최적화 - GQA(Grouped Query Attention) 도입	8K	Mobile/Chatbot
Qwen 2.5-Base (2024.12)	1.5B~72B	- 다국어(119개 언어) 데이터셋 확장 - RAG 최적화 및 긴 문맥 처리 능력 강화	32K~256K	다국어, RAG, 코드
Qwen 2.5-Max (2025.01)	72B(MoE)	- MoE(Mixture-of-Experts) 아키텍처 도입 (64 Experts) - 100만 토큰 장문 처리 및 논리 추론 강화	1M	멀티모달, 장문, 산업

151) NDTV World, 2025, Alibaba Qwen2.5 Max beats rivals DeepSeek & GPT-4o, https://ndtv.com/world-news/alibaba-qwen-beats-deepseek-gpt4o
152) LMSYS Org, 2025, Chatbot Arena Leaderboard: Coding & Hard Prompts - Qwen Ranking, https://chat.lmsys.org
153) Alibaba Cloud Blog, 2025, Alibaba Launches Qwen App to Boost its Consumer AI Efforts, https://www.alibabacloud.com/blog/602672

Qwen VL (2025.09)	-	- Native Multimodal (비전・오디오 통합 학습) - 초장문(2M) 컨텍스트 및 실시간 스트리밍 처리	2M	비전・오디오・추론

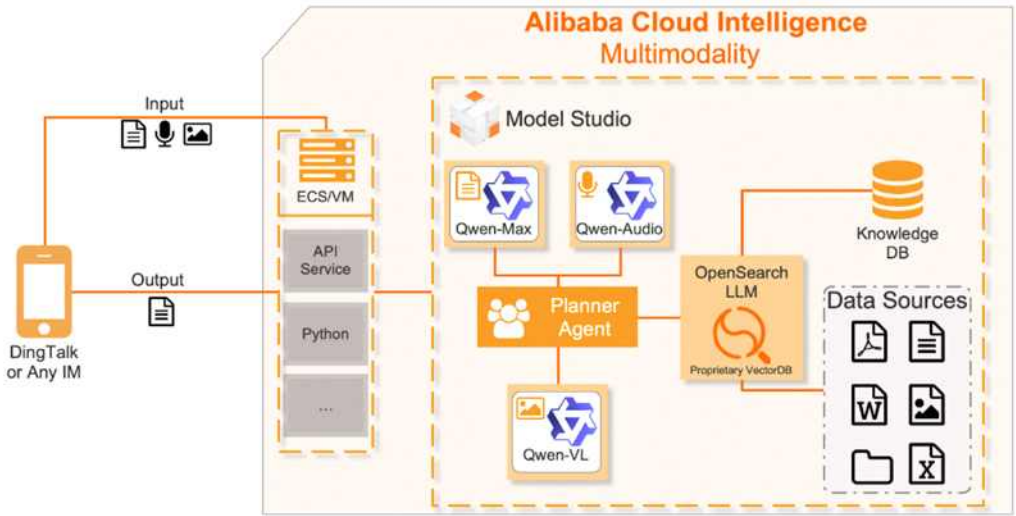
(자체 작성)

Qwen 시리즈는 단순한 텍스트 생성을 넘어 코딩 어시스턴트, 멀티모달(Vision, Audio), 구조화 데이터 분석 등 다양한 기능을 내재화한 "올인원(All-in-One)" 플랫폼으로 진화하였다.

1) 아키텍처 효율화 및 성능 고도화[154]

최신 Qwen 2.5-Max는 64개의 전문가(Expert) 네트워크로 구성된 72B MoE 구조를 채택하였다. 동적 비활성화 기술을 통해 연산 비용을 기존 대비 30% 절감하면서도, 20조 토큰 이상의 방대한 학습량과 50만 건 이상의 인간 평가(RLHF)를 통해 추론 능력을 극대화하였다. 그 결과, GSM8K(수학) 94.5점, HumanEval(코딩) 81.5점 등 주요 벤치마크에서 SOTA(State-of-the-Art)를 기록하였다.

[그림 23] 알리바바 큐웬 멀티모달 통합 아키텍처



출처 : https://www.alibabacloud.com/blog/building-multimodal-services-with-qwen-and-model-studio_600962 High-Level Architecture Overview)

154) NDTV World, 2025, Alibaba Qwen2.5 Max beats rivals DeepSeek & GPT-4o, https://ndtv.com/world-news/alibaba-qwen-beats-deepseek-gpt4o

2) 엔터프라이즈 최적화 및 확장성[155]

Qwen은 100만 토큰(1M+) 이상의 컨텍스트 윈도우를 지원하여, Llama 대비 약 8배 긴 문맥 처리 능력을 자랑한다. 이는 금융, 법률, 헬스케어 등 대규모 문서를 다루는 엔터프라이즈 환경에 최적화된 강점이다. 또한 Flink, ONNX 등 다양한 런타임 환경을 지원하고 양자화(Quantization) 및 LoRA 등 경량화 기술을 제공하여, 클라우드부터 엣지 디바이스까지 유연한 배포가 가능하다.

<표 9> 알리바바 큐웬 개요

<table>
<tr><td>개발자</td><td colspan="3">알리바바(Alibaba Cloud)</td><td>홈페이지</td><td>https://qwen.ai</td></tr>
<tr><td>최초공개 시기</td><td colspan="3">2023.04</td><td>라이선스</td><td>Apache 2.0
(Qwen 2.5 기준,
학습/추론/상업 완전 자유,
OSI 공식 충족)</td></tr>
<tr><td>주요 특징</td><td colspan="5">글로벌 오픈소스 리더 / 대형 MoE / 멀티모달 / 119개언어 / 산업・학술・API 확대</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/qwen</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>157</td><td>24</td><td>395</td><td colspan="2">5</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Qwen2.5-Max</td><td>358,120</td><td>1,870</td><td>21,451</td><td>88</td></tr>
<tr><td rowspan="4">주요 활용 사례</td><td colspan="3">(FAW Group) 자동차 제조, 정책분석, 생산관리</td><td colspan="2">https://www.alibabacloud.com/blog/qwen-ecosystem-expands-rapidly-accelerating-ai-adoption-across-industries_602330</td></tr>
<tr><td colspan="3">(Ant Financia): 금융 데이터 분석, 문서 자동화</td><td colspan="2">https://www.alibabacloud.com/blog/qwen-ecosystem-expands-rapidly-accelerating-ai-adoption-across-industries_602330</td></tr>
<tr><td colspan="3">(공식 Qwen 앱) 챗봇・AI 에이전트, 1,000만+ 다운로드 (개인・비즈니스)</td><td colspan="2">https://qwen.ai/app</td></tr>
<tr><td colspan="3">Ping An Healthcare: 의료 챗봇, 임상보고,</td><td colspan="2">https://www.alibabacloud.com/blog/qwen-ec</td></tr>
</table>

155) DataCamp, 2025, Qwen 2.5-Max: Features, DeepSeek V3 Comparison & More, https://www.datacamp.com/blog/qwen-2-5-max

	문서생성	osystem-expands-rapidly-accelerating-ai-adoption-across-industries_602330
	(Univa) IT 개발 코드 자동화, 오픈소스 파생	https://github.com/QwenLM/Qwen-1_8B
기술 데이터 공개 범위 및 주요 라이선스 가중치 공개 : O (HuggingFace, Github 즉시 다운로드, 클라우드 배포) 학습/추론 코드 공개 : O (공식 Github, PyTorch・Transformers・vLLM 등 완전 통합) 데이터 공개 : 부분 공개(일부 자체・공개셋), 나머지 “publicly available sources“ 명시 라이선스 유형 : Apache 2.0 (상업・비상업・연구 무제한, Acceptable Use 없음) OSI 기준 충족 : O (공식 선언, EU/NA/아시아 전역 자유) 상업적 제한 : 없음, B2B/B2C 클라우드・API・온프레미스 모두 개방		

(자체 작성)

3. 딥시크

DeepSeek(High-Flyer Quantum, DeepSeek AI)는 “효율성의 끝판왕“으로 불리는 중국의 오픈소스 AI 모델 시리즈다. GPT-4o 및 Claude-3.5와 대등한 성능[156]을 내는 DeepSeek-V3(671B MoE)를 단 560만 달러(약 80억 원)라는 충격적인 저비용으로 학습시켜, 글로벌 AI 업계에 “The DeepSeek Shock” [157]를 일으켰다. 특히 자체 개발한 MLA(Multi-Head Latent Attention)와 DeepSeekMoE 아키텍처를 통해 연산 효율을 극한으로 끌어올렸으며, 이를 바탕으로 API 가격[158]을 경쟁사 대비 1/10~1/100 수준으로 낮춰 ’AI 가격 전쟁’을 촉발하였다.

2025년 초 DeepSeek-R1(Reasoning 특화 모델)을 통해 OpenAI o1 수준의 논리 추론 능력을 오픈소스로 공개하며 기술적 리더십을 입증했으나, 한국 등 일부 국가에서는 개인정보 유출 우려로 서비스가 일시 차단되는 등 보안 논란[159]도 겪고 있다.

이런 논란에도 불구하고, 중국 AI 경쟁력을 대변하는 모델로써 출시 1개월 만에 다양한 산업에 적용되는 표준 AI 모델로 자리를 잡았으며, 이는 “AI가 실험실

156) Galileo, 2025, DeepSeek R1 vs OpenAI o1 Comparison, https://galileo.ai/blog/deepseek-r1-vs-openai-o1-comparison
157) DeepSeek-AI, 2024, DeepSeek-V3 Technical Report: Mixture-of-Experts & MLA Architecture, https://arxiv.org/abs/2412.19437
158) DeepSeek-AI, 2024, DeepSeek-V3 Technical Report: Mixture-of-Experts & MLA Architecture, https://arxiv.org/abs/2412.19437
159) Reuters, 2025, South Korean ministries block DeepSeek on security concerns, https://www.reuters.com/technology/artificial-intelligence/south-koreas-industry-ministry-temporarily-bans-access-deepseek-security-concerns-2025-02-05/

을 벗어나 제조/하드웨어(On-Device)와 결합될 때, '저비용 고지능'이 핵심 트리거[160]임을 증명하기도 하였다.

<표 10> 딥시크 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
DeepSeek-Coder-V1 (2023.11)	33B (Dense)	- 최초 공개 모델 - Code Llama 기반 개선 - 2T 토큰 코드 데이터 학습	16K	[초기 진입/코딩] 단순 코드 자동완성
DeepSeek-V2 (2024.05)	236B / 21B (MoE)	- MLA (Multi-head Latent Attention) 최초 도입 - DeepSeekMoE 아키텍처	128K	[효율성 혁신] API 서비스, 대규모 텍스트 처리
DeepSeek-V3 (2024.12)	671B / 37B (MoE)	- MoE 완성형 - FP8 Mixed Precision 학습 - Aux-Loss-Free 로드 밸런싱	128K	[플래그십/범용] GPT-4o 대체, 엔터프라이즈 표준
DeepSeek-R1 (2025.01.20)	671B / 37B (MoE)	- Pure RL (강화학습) - GRPO (Group Relative Policy) - CoT(사고과정) 내재화	128K	[추론/Reasoning] 수학, 과학, 복잡한 문제 해결
DeepSeek-V3.1 (2025.08.21)	671B 기반 (Hybrid)	- Thinking / Non-thinking 통합 - R1(추론)과 V3(챗) 능력 결합 - 코딩/터미널 사용 능력 대폭 강화	128K	[올인원/하이브리드] 모드 전환 없이 일반 대화+심층 추론
DeepSeek-V3.2-Exp (2025.09.29)	미공개 (Sparse)	- Huawei Ascend (CANN) 최적화 - DeepSeek Sparse Attention - Non-NVIDIA 생태계 첫 지원	128K	[소버린/내수용] 중국 내수 인프라, 저비용 장문 처리

(자체 작성)

(1) MLA (Multi-Head Latent Attention) - "메모리 혁명"[161]

기존 MHA(Multi-Head Attention)는 모델이 커질수록 KV Cache(이전 대화 기억 메모리)가 기하급수적으로 늘어나는 병목이 있었다. DeepSeek는 이를 해결하기 위해 KV를 저차원 잠재 벡터(Latent Vector)로 압축하여 저장하고, 필요할 때만 복원하는 기술을 도입[162]하였다. 이로 인해 추론 시 메모리 점유율을 최대 93%까

160) CNN, 2025, A shocking Chinese AI advancement called DeepSeek is sending US stocks plunging, https://www.cnn.com/2025/01/27/tech/deepseek-stocks-ai-china

161) Vizuara, 2025, Decoding Multi-Head Latent Attention (MLA): Reducing KV Cache by 93%, https://vizuara.substack.com/p/decoding-multi-head-latent-attention

지 줄였다.

이러한 기술적용은 “성능 vs 효율“의 오랜 트레이드 오프를 깼다. 단일 노드(8개 GPU)에서도 671B급 초대형 모델을 고속으로 돌릴 수 있게 되면서, API 가격을 경쟁사 대비 1/20로 낮추는 결정적 원동력[163]이 되었다. 이는 메타(Llama)와 Google이 차세대 모델 설계 시 ’효율화’에 사활을 걸게 만든 기술적 트리거가 되었다고 평가한다.

(2) DeepSeekMoE (Fine-Grained Experts) - “전문가의 세분화” [164]

기존 MoE(예: Mixtral)가 소수의 큰 전문가(Experts) 8명 중 2명을 골라 썼다면, DeepSeek는 전문가를 매우 잘게 쪼개고, 일부는 공유 전문가(Shared Expert)로 고정하여 항상 활성화하였다. (총 671B 파라미터 중 토큰당 단 37B(약 5.5%)만 활성)

결과적으로 “전문가 특화“와 “공통 지식 유지“를 동시에 달성하였다. 지식이 파편화되는 MoE의 단점을 보완하면서도, 모델 사이즈 대비 실제 구동 비용은 1/10 수준으로 유지한 것이다. 이는 “거대 모델의 상품화(Commodification)“ 시기를 1년 이상 앞당긴 혁신으로 평가받고 있다.

(3) FP8 Mixed Precision Training - “하드웨어의 한계 극복” [165]

학습의 전 과정(Forward/Backward)을 FP8(8비트 부동소수점) 정밀도로 수행하는 프레임워크를 자체 구축하여, H800 GPU 클러스터의 연산 효율을 극한까지 끌어올렸다. 최신 H100을 구하기 어려운 상황에서도 알고리즘 최적화만으로 GPT-4급 모델을 만들 수 있음을 증명하여, 글로벌 AI 패권 경쟁의 양상을 ’자원 싸움’에서 ’기술 효율성 싸움’으로 전환시[166]켰다.

---

162) Towards Data Science, 2025, DeepSeek-V3 Explained: Multi-Head Latent Attention, https://towardsdatascience.com/deepseek-v3-explained-1-multi-head-latent-attention-ed6bee2a67c4
163) DeepSeek-AI, 2024, DeepSeek-V3 Technical Report, https://arxiv.org/abs/2412.19437
164) Creative Strategies, 2025, DeepSeek MoE & V2: Commoditizing Large Models, https://creativestrategies.com/deepseek-moe-v2/
165) LinkedIn (Qi He), 2025, DeepSeek V3's Key Innovations in 8-bit Floating Point (FP8) Learning, https://www.linkedin.com/pulse/deepseek-v3s-key-innovations-8-bit-floating-point-fp8-qi-he-e3dgf
166) AInvest, 2025, DeepSeek Blows Up 메타's AI Strategy: A Paradigm Shift in the AI Race, https://www.ainvest.com/news/deepseek-blows-메타-ai-strategy-paradigm-shift-ai-race-2504/

<표 11> 딥시크 개요

<table>
<tr><td>개발자</td><td colspan="2">Deepseek</td><td>홈페이지</td><td>https://www.deepseek.com/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2023.11 (Coder), 2024.12 (V3)</td><td>라이선스</td><td>MIT License (완전 오픈소스, 상업적 이용 자유)</td></tr>
<tr><td>주요 특징</td><td colspan="4">가격 파괴자(GPT-4급 성능, 1/20 비용) / MLA & DeepSeekMoE 아키텍처 혁신 / 추론(R1) 및 코딩 특화</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="4">https://huggingface.co/deepseek-ai</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td>데이터셋 수</td></tr>
<tr><td>31</td><td>16</td><td>82</td><td>2</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>DeepSeek-V3</td><td>420,000,000+</td><td>12,800+</td><td>21,450+</td><td>350+</td></tr>
<tr><td rowspan="5">주요 활용 사례</td><td colspan="2">중국 전기차(EV) 업계의 ’지능형 비서(Smart Cockpit)’ 표준 모델로 활용</td><td colspan="2">https://www.chinadailyhk.com/hk/article/604736</td></tr>
<tr><td colspan="2">내부 개발 도구 및 대고객 챗봇 서비스의 기반 모델로 DeepSeek 아키텍처 채택
*SMB(중소기업) 고객 응대. 월 $10 미만 비용으로 24시간 주문/예약 처리 봇 구축. (기존 GPT-4 대비 1/20 비용)</td><td colspan="2">https://www.byteplus.com/en/topic/416739?title=deepseek-chatbot-whatsapp-2025</td></tr>
<tr><td colspan="2">WeChat / Lark (Feishu) - 백엔드 LLM 전환. 텐센트와 바이트댄스가 자사 협업 툴의 AI 요약/번역 엔진을 자체 모델에서 DeepSeek 기반으로 전환</td><td colspan="2">https://www.cnbc.com/2025/02/21/deepseek-led-ai-adoption-offers-china-opportunity-to-boost-growth.html</td></tr>
<tr><td colspan="2">대출 심사 & 약관 분석. 골드만삭스 보고서에 따르면, 중국 금융권의 사무 자동화 도입으로 GDP 생산성 0.2%p 향상 예상.</td><td colspan="2">https://www.morganstanley.com/insights/articles/deepseek-ai-watershed-moment</td></tr>
<tr><td colspan="2">차이나모바일 - 장애 원인 분석 (RCA). 통신 장애 로그를 DeepSeek-Coder로 분석하여 원인 파악 시간 50% 단축.</td><td colspan="2">https://chat-deep.ai/news/deepseek-ai-industrial-integrations-telecom-auto-2025/</td></tr>
<tr><td colspan="5">기술 데이터 공개 범위 및 주요 라이선스
가중치 공개 : O (완전 공개) (HuggingFace, ModelScope 등 즉시 다운로드)
코드 공개 : O (완전 공개) (학습/추론/평가 코드 포함 GitHub 공개)
데이터 공개 : 부분 공개 (데이터셋 구성 비율 및 전처리 방식 논문 공개, 원본 데이터 미공개)
라이선스 : MIT License (가장 자유로운 오픈소스 라이선스 중 하나)
상업적 제한 : 없음 (누구나 무료로 상업적 서비스/제품 구축 가능)
OSI 기준 : O (충족) (Llama와 달리 제약 조건이 거의 없어 진정한 오픈소스 인정)</td></tr>
</table>

(자체 작성)

4. 미스트랄AI

Mistral AI는 프랑스 파리에 본사를 둔 유럽 대표 AI 스타트업으로, “오픈 가중치(Open-Weight) 기반의 고효율 모델“을 지향한다.

2025년 기준 기업가치 117억 유로(약 17조 원)를 인정[167]받았으며, ASML・Microsoft・Nvidia 등으로부터 대규모 투자를 유치[168]하였다.

Mistral의 핵심 철학은 “크기는 작게, 지능은 높게(Small but Mighty)” 다. 수천억 파라미터 경쟁 대신 Sliding Window Attention(SWA), Sparse Mixture-of-Experts(SMoE) 등 아키텍처 최적화를 통해 하드웨어 요구사항을 획기적으로 낮췄다.

특히 2025년 출시된 Magistral(추론 특화)와 Pixtral(멀티모달)을 통해 텍스트를 넘어 복합 지능 영역으로 확장했으며, “소버린(Sovereign) AI” 전략을 통해 데이터 주권을 중시하는 유럽 기업 및 공공기관의 표준 모델[169]로 자리 잡았다.

<표 12> 미스트랄AI 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
Mistral 7B(2023.09)	7B	-SWA (Sliding Window Attention) 최초 도입- GQA (Grouped-Query Attention) 적용	32K	[입문/범용]개인용 챗봇, 텍스트 요약
Mixtral 8x7B(2023.12)	47B / 13B(MoE)	-최초의 고성능 오픈소스 MoE- 토큰당 2개 전문가 활성화 (Top-2 Routing)	32K	[가성비 고성능]GPT-3.5 대체, 다국어 번역
Mistral Large 2(2024.07)	123B	-플래그십 Dense 모델- 다국어 (유럽 5개국어) 최적화- 코딩/수학 성능 대폭 강화	128K	[엔터프라이즈]복합 업무, RAG, 정밀 추론
Mistral Small	24B	-온디바이스/엣지 최적화- 150	32K	[실시간/엣지]모바일 비

167) BeBeez, 2025, Europe's AI bet: Paris-based Mistral wins €1.7 billion, doubling valuation to €11.7 billion, https://bebeez.eu/2025/09/09/europes-ai-bet-paris-based-mistral-wins-e1-7-billion-doubling-valuation-to-e11-7-billion/

168) CNBC, 2025, AI firm Mistral valued at $14 billion as chip giant ASML takes major stake, https://www.cnbc.com/2025/09/09/ai-firm-mistral-valued-at-14-billion-as-asml-takes-major-stake.html

169) AI Competence, 2025, Mistral AI: Europe’s Bold Move For AI Sovereignty, https://aicompetence.org/mistral-ai-europes-bold-move-for-ai-sovereignty/

3.1(2025.03)		token/s 초고속 추론		서, 로컬 디바이스 제어
Magistral (2025.07)	비공개	-Reasoning (System 2)- CoT 강화, 강화학습(RL) 파이프라인	128K	[심층 추론]전략 기획, 과학 연구, 복잡 코딩

(자체 작성)

(1) SMoE (Sparse Mixture-of-Experts) - "MoE 대중화의 시초 "[170]

Mixtral 8x7B를 통해 MoE 아키텍처가 오픈소스 진영에서도 GPT-3.5급 성능을 낼 수 있음을 최초로 증명하였다. 총 47B 파라미터 중 토큰당 단 13B만 활성화하여 연산 효율을 극대화하였다.

이는 "모델은 무조건 커야 한다"는 고정관념을 깼다. 활성 파라미터(Active Parameter)를 최소화하여 추론 비용을 1/6(Llama 2 70B 대비)로 줄이는 효율성 혁명을 주도했으며, 이후 DeepSeek와 Qwen이 MoE를 채택하는 데 결정적인 영감을 주었다.

(2) Magistral (System 2 Thinking) - "유럽 추론 모델"[171][172]

2025년 7월 출시된 Mistral의 첫 번째 추론(Reasoning) 모델이다. DeepSeek-R1이나 OpenAI o1처럼 "생각하는 시간(Test-time Compute)"[173]을 활용해 복잡한 수학, 코딩, 전략 기획 문제를 해결한다. 단순 텍스트 생성을 넘어 "의사결정 및 기획(Planning)" 영역으로 오픈소스 AI의 활용 범위를 확장하였다. 특히 유럽 내 금융/연구 기관에서 데이터 유출 걱정 없이 고난도 추론을 수행할 수 있는 유일한 대안이 되었다. 이를 기반으로 유럽 데이터 주권 수호를 위한 대표 모델로 인식되고 있다.

2025년 VivaTech에서 발표된 전략으로, NVIDIA H100/Blackwell B200 기반의 유럽 내 자체 슈퍼컴퓨팅 인프라를 구축중이다. 데이터가 미국 서버로 넘어가지 않도록 물리적으로 보장하였으며 GDPR 및 EU AI 법안(EU AI Act) 준수가 필수적

170) arXiv, 2024, Mixtral of Experts, https://arxiv.org/abs/2401.04088
171) BrainIllustrate, 2025, Mistral AI: The Open-Core Challenger Forging a New Path, https://www.brainillustrate.com/2025/09/mistral-ai-open-core-challenger-forging.html
172) Aivancity, 2025, VivaTech 2025: Mistral AI unveils sovereign HPC infrastructure, https://www.aivancity.ai/blog/vivatech-2025-mistral-ai-unveils-a-sovereign-hpc-infrastructure-in-partnership-with-nvidia/
173) Weights & Biases (W&B), 2025, Mistral AI Debuts Magistral: A Reasoning-Centric Language Model, https://wandb.ai/byyoung3/ml-news/reports/Mistral-AI-Debuts-Magistral-a-Reasoning-Centric-Language-Model--VmlldzoxMTEyMjUyNg

인 유럽 금융, 국방, 공공기관 시장을 독점적으로 확보하는 기반이 되었다.

<표 13> 미스트랄AI 개요

<table>
<tr><td>개발자</td><td colspan="2">Mistral AI</td><td>홈페이지</td><td colspan="2">https://mistral.ai/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2023.09 (Mistral 7B) / 2025.07 (Magistral)</td><td>라이선스</td><td colspan="2">Apache 2.0 (7B, 8x7B 등 소형), Mistral Commercial (Large, Magistral)</td></tr>
<tr><td>주요 특징</td><td colspan="5">유럽 대표 소버린 AI / 고효율 소형 모델(SWA, SMoE) / 데이터 주권 및 기업 보안 특화</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/mistralai</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>42</td><td>14</td><td>38</td><td colspan="2">3</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Mistral-7B-v0.1</td><td>38,500,000+</td><td>14,500+</td><td>N/A</td><td>200+</td></tr>
<tr><td rowspan="4">주요 활용 사례</td><td colspan="2">Stellantis (스텔란티스) - 전사적 AI 도입. 차량 내 음성 비서뿐만 아니라 영업/엔지니어링 워크플로우 전체에 Mistral 모델 통합</td><td colspan="3">https://www.stellantis.com/en/news/press-releases/2025/october/stellantis-and-mistral-ai-expand-their-collaboration-to-accelerate-enterprise-wide-ai-adoption</td></tr>
<tr><td colspan="2">ASML-반도체 R&D 가속화. 13억 유로 투자와 함께 리소그래피 시스템 개발 및 운영 효율화에 Magistral 모델 활용</td><td colspan="3">https://research.contrary.com/company/mistral-ai</td></tr>
<tr><td colspan="2">TotalEnergies - 에너지 데이터 분석. 내부 구축형(On-Premise) LLM으로 보안 규제를 준수하며 에너지 탐사 데이터 분석</td><td colspan="3">https://totalenergies.com/news/press-releases/totalenergies-collaborate-mistral-ai-increase-application-artificial</td></tr>
<tr><td colspan="2">Capgemini-글로벌 컨설팅. 전 세계 기업 고객에게 Mistral 기반의 생성형 AI 솔루션 구축 및 최적화 서비스 제공</td><td colspan="3">https://www.capgemini.com/about-us/technology-partners/mistral-ai/</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스 (Mistral Small 3.1 / Mixtral 기준)
* Mistral Large 2 / Magistral은 클로즈드 모델
가중치 공개 : O (완전 공개)
코드 공개 : O (완전 공개) (참조 코드 포함)
데이터 공개 : X (미공개)
라이선스 : Apache 2.0 (상업적 이용 자유)
상업적 제한 : 없음
OSI 기준 : O (충족)</td></tr>
</table>

(자체 작성)

5. 구글 Gemma/BERT

Gemma 3는 구글 DeepMind가 개발한 최신 경량 오픈 모델(Open Model) 시리즈로, Gemini 2.0의 연구 기술을 기반으로 구축되었다. 2025년 3월 출시된 Gemma 3는 기존 텍스트 중심 모델에서 벗어나 아키텍처 단계부터 이미지와 텍스트를 동시에 이해하는 멀티모달 네이티브(Multimodal Native) 능력을 탑재했으며, 128K 토큰의 장문 컨텍스트를 지원한다. 핵심 전략은 “책임감 있는 AI(Responsible AI)“와 “접근성(Accessibility)“으로, 개발자가 일반 노트북이나 모바일 기기(온디바이스)에서 직접 구동할 수 있는 1B~27B 사이즈의 고효율 모델에 집중하고 있다.

Google BERT는 2018년 발표 이후 ’자연어 이해(NLU)’의 표준으로 자리 잡았으나, 2024년 12월 ModernBERT로 재탄생하며 제2의 전성기를 맞았다. 기존 BERT의 한계(512 토큰 제한)를 극복하고 8,192 토큰의 장문 처리를 지원하며, 최신 Flash Attention 2 기술로 추론 속도를 비약적으로 높여 2025년 현재 RAG 시스템의 핵심 리트리버(Retriever)로 재조명받고 있다.

<표 14> 구글 주요 오픈소스 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
Gemma 2(2024.06)	9B / 27B	-Knowledge Distillation적용- Logit Soft-capping (학습 안정화)	8K	[고성능 경량]추론, 요약, 챗봇 백엔드
Gemma 3 1B(2025.03)	1B(MoE)	-멀티모달 (Vision) 탑재- 초경량 MoE 아키텍처- 모바일 NPU 최적화	128K	[초경량/모바일]IoT 센서 분석, 스마트폰 비서
Gemma 3 4B(2025.03)	4B(MoE)	-온디바이스 표준- 이미지 캡셔닝/VQA 가능- 엣지 디바이스용 최적 밸런스	128K	[온디바이스]노트북 AI, 엣지 챗봇, 문서 요약
Gemma 3 27B(2025.03)	27B	-연구/서버용 고성능- 복합 추론 및 정밀 RAG- Llama 70B급 성능 지향	128K	[연구/서버]복잡한 지시 이행, 코드 생성
BERT-Base(2018.10)	110M	-Encoder-Only Transformer- MLM (Masked Language Model)- NSP (Next Sentence Prediction)	512	[일반 NLU]텍스트 분류, 개체명 인식(NER), 감성 분석

DistilBERT(2019.10)	66M	-Knowledge Distillation(BERT 경량화)- 레이어 수 50% 축소, 속도 60% 향상- 성능 97% 유지	512	[온디바이스/실시간]모바일 앱 내 검색, 실시간 댓글 필터링
RoBERTa-Large(2019.07)	355M	-Dynamic Masking(학습 효율 개선)- NSP 제거, 더 많은 데이터/시간 학습- BERT 대비 강건성(Robustness) 강화	512	[고성능 NLU]정밀 문서 분류, 팩트 체크, 모순 탐지
DeBERTa-V3(2021.06)	400M	-Disentangled Attention(위치/내용 분리)- Enhanced Mask Decoder- 기존 BERT 계열 중 SOTA 성능	512	[SOTA급 NLU]복잡한 질의응답(QA), 의미론적 텍스트 유사도
ModernBERT(2024.12)	110M / 340M	-Rotary Embeddings (RoPE)(장문 처리)-Flash Attention 2(속도 2배↑)- Unpadding(토큰 효율화)	8,192	[차세대 검색/RAG]RAG 리트리버, 장문 법률/금융 문서 분석

(자체 작성)

(1) Gemma 3 : 멀티모달 네이티브 (Native Multimodal) - “보는 AI의 대중화” 174)

Gemma 3는 별도의 비전 인코더를 부착하는 방식이 아니라, 모델 아키텍처 자체에서 이미지 토큰과 텍스트 토큰을 동등하게 처리하도록 설계되었다. 이로 인해 1B급 초소형 모델에서도 정교한 이미지 캡셔닝과 시각적 질의응답(VQA)이 가능하다. 이는 스마트폰이나 IoT 기기에서 서버 연결 없이 카메라 입력을 실시간으로 분석할 수 있게 하여 엣지 컴퓨팅(Edge Computing) 시장에 큰 파장을 일으켰다.

(2) Gemma 3: Sliding Window & Local Attention - “장문 처리의 확산” 175)

128K라는 긴 컨텍스트를 효율적으로 처리하기 위해, 모든 레이어가 전체 문맥을 보는 대신 일부 레이어(Local Layer)는 1024 토큰 범위만 집중하도록 설계하였

174) Google DeepMind, 2025, Gemma 3 Technical Report, https://arxiv.org/abs/2503.19786
175) Google Developers Blog, 2025, Gemma explained: What's new in Gemma 3, https://developers.googleblog.com/en/gemma-explained-whats-new-in-gemma-3/

다. 이를 통해 긴 문서를 분석할 때도 KV Cache 메모리 사용량을 획기적으로 줄여, 고가의 서버용 GPU가 아닌 RTX 3090/4090급 일반 소비자용 GPU에서도 원활한 장문 분석이 가능[176]해졌다.

(3) ModernBERT: 2025년형 리팩토링 (The Comeback) - “검색의 질을 바꾸다” [177]

2018년의 낡은 BERT 구조를 2025년 최신 기술로 완벽하게 리팩토링[178]하였다. Rotary Embeddings (RoPE)와 Unpadding 기술을 적용해 고질적인 512 토큰 제한을 8,192 토큰으로 확장했으며, Flash Attention 2를 통해 추론 속도를 2~4배 높였다. 이로 인해 ModernBERT는 생성형 AI(LLM)의 할루시네이션을 방지하는 RAG 시스템에서, “어떤 문서를 참고할지“ 결정하는 리트리버(Retriever) 모델의 새로운 표준이 되었다.

(4) 도메인 특화 (Domain Adaptation) - “기업 내부의 스페셜리스트” [179]

BERT 계열은 LLM 대비 파인튜닝 비용이 1/100 수준으로 매우 저렴하다. 이러한 장점 덕분에 BioBERT(의생명), FinBERT(금융), LegalBERT(법률) 등 특정 분야 데이터로 학습된 특화 모델들이 전 세계 병원, 은행, 로펌의 내부 시스템(On-Premise)을 장악하고 있다. 데이터 보안 문제로 클라우드 LLM 도입을 꺼리는 기업들에게 가장 현실적인 AI 솔루션으로 평가받는다.

<표 15> 구글 젬마(Gemma) 개요

개발자	Google DeepMind	홈페이지	https://ai.google.dev/gemma
최초공개 시기	2024.02 (Gemma 1) / 2025.03 (Gemma 3)	라이선스	Gemma Terms of Use (상업적 이용 허용, 책임 있는 AI 사용 조건)

176) Rohan Paul (AI Researcher), 2025, Google released Gemma 3: 128k Long-Context Window, https://www.rohan-paul.com/p/google-released-gemma-3-128k-long
177) Hugging Face Blog, 2025, Finally, a Replacement for BERT: Introducing ModernBERT, https://huggingface.co/blog/modernbert
178) arXiv, 2024, Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, https://arxiv.org/abs/2412.13663
179) Jina AI, 2025, What should we learn from ModernBERT?, https://jina.ai/news/what-should-we-learn-from-modernbert/

<table>
<tr><td>주요 특징</td><td colspan="5">온디바이스 & 멀티모달 최적화 / Gemini 기술 기반 오픈 모델 / Android 및 Google 생태계(Kaggle, Colab) 연동</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/google</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>3348</td><td>46</td><td>1066</td><td colspan="2">63</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Gemma-2-9b-it</td><td>11,500,000+</td><td>8,500+</td><td>N/A</td><td>150+</td></tr>
<tr><td rowspan="3">주요 활용 사례</td><td colspan="3">Android / Pixel - 온디바이스 AI 기능.
삼성 갤럭시 및 구글 픽셀 폰의 로컬 AI 기능(실시간 번역, 요약) 구현에 경량화된 Gemma 모델 활용.</td><td colspan="2">https://io.google/2025/explore/pa-keynote-4</td></tr>
<tr><td colspan="3">NVIDIA NIM-엔터프라이즈 배포. NVIDIA의 최적화된 추론 마이크로서비스(NIM)를 통해 기업들이 Gemma 모델을 온프레미스/클라우드에 쉽게 배포</td><td colspan="2">https://catalog.ngc.nvidia.com/orgs/nim/teams/google/containers/gemma-3-1b-it</td></tr>
<tr><td colspan="3">Kaggle / Colab- 데이터 사이언스/교육.
전 세계 데이터 과학자들이 Gemma를 활용해 경진대회 참여 및 AI 모델 튜닝 실습. 가장 접근성 높은 모델.</td><td colspan="2">https://ai.google.dev/gemma/docs/core?hl=ko</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스
가중치 공개 : O (완전 공개) (Kaggle, HuggingFace, Vertex AI Model Garden)
코드 공개 : O (완전 공개) (Inference, Fine-tuning 예제 코드 포함)
데이터 공개 : X (미공개) (학습 데이터 구성 비율 등 상세 미공개, 안전 필터링 강조)
라이선스 : Gemma Terms of Use (개방형 라이선스이나 Apache 2.0 보다는 제약 있음)
상업적 제한 : 없음 (단, 유해한 용도 금지 등 AUP 준수 필수)
OSI 기준 : X (불충족) (엄밀한 의미의 오픈소스 정의와는 다름, 'Open Model' 표방)</td></tr>
</table>
(자체 작성)

6. OpenAI gpt-oss

GPT-OSS는 그동안 폐쇄형(Closed) 정책을 고수해 온 OpenAI가 2025년 8월, 전략을 수정하여 최초로 공개한 공식 오픈 웨이트(Open-Weight) 모델 시리즈[180]다. DeepSeek 등 고성능 오픈소스 모델의 약진에 대응하기 위해 출시되었으며, 기존 'o 시리즈(o1, o3)'의 강력한 추론(Reasoning) 능력을 경량화된 오픈 모델로 이식하는 데 성공하였다.

핵심 모델인 gpt-oss-120b와 gpt-oss-20b는 Apache 2.0 라이선스를 채택하여 연

180) OpenAI, 2025, Introducing gpt-oss: Pushing the frontier of open-weight reasoning models, https://openai.com/index/introducing-gpt-oss/

구뿐만 아니라 상업적 이용과 수정이 완전히 자유롭다. 특히 MXFP4(4-bit) 양자화 기술과 MoE(Mixture-of-Experts) 아키텍처를 결합하여, H100 단일 GPU에서도 120B급 모델을 고속으로 구동할 수 있는 효율성을 갖췄다. 이는 “OpenAI 성능의 대중적 확산”을 상징하는 모델로 평가받는다.[181]

<표 16> OpenAI 주요 오픈소스 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
gpt-oss-120b (2025.08)	117B	- 추론 최적화 MoE (Reasoning-focused MoE) - MXFP4 네이티브 양자화 (H100 단일 구동) - Apache 2.0 라이선스 (상업적 완전 자유)	128K	- 엔터프라이즈 내부망 (On-Premise) 추론 - 복잡한 자율 에이전트 개발 - 고성능 코딩 및 수학 문제 해결
gpt-oss-20b (2025.08)	21B	- 고밀도 트랜스포머 아키텍처 - MXFP4 양자화 (RTX 4090 구동 가능) - Apache 2.0 라이선스	32K	- 로컬 코딩 어시스턴트 - 엣지(Edge) 디바이스 추론 - 개인 연구자 및 개발자

(자체 작성)

(1) 고효율 MoE 아키텍처 (Efficient MoE) - “120B의 성능, 5B의 가벼움”[182]

GPT-OSS는 120B 모델 기준으로 총 117B 파라미터 중 토큰당 단 5.1B(약 4.3%)만 활성화하는 초고효율 MoE(Mixture-of-Experts) 구조를 채택하였다. 이는 DeepSeek나 Mixtral보다 더 극단적인 희소성(Sparsity)을 구현한 것으로, Softmax-after-topk 라우팅 알고리즘을 통해 전문가 선택의 정확도를 높였다. 결과적으로 120B급 지능을 유지하면서도 실제 연산량은 7B 모델 수준으로 낮춰, 추론 속도를 비약적으로 향상시켰다.

(2) Native MXFP4 양자화 - “H100 한 장의 혁명”[183]

181) TechXplore, 2025, OpenAI chief says it needs new open-source strategy, https://techxplore.com/news/2025-02-openai-chief-source-strategy.html
182) OpenAI, 2025, Introducing gpt-oss: Pushing the frontier of open-weight reasoning models, https://openai.com/index/introducing-gpt-oss/

기존의 사후 양자화(Post-Training Quantization)와 달리, 학습 단계부터 4비트 포맷인 MXFP4 (Micro-scaling formats 4-bit)를 네이티브로 적용하였다. 특히 전체 파라미터의 90%를 차지하는 MoE 레이어 가중치를 4비트로 정밀하게 압축하여, 메모리 사용량을 1/4로 줄였다. 이 기술 덕분에 단일 H100 (80GB) GPU에서도 120B 모델을 지연 없이 구동할 수 있으며, 20B 모델은 16GB 메모리를 가진 소비자용 GPU(RTX 4080/4090)에서도 원활히 작동한다.

(3) Chain-of-Thought (CoT) 통합 추론 - “생각하는 오픈 모델” 184)

GPT-OSS는 o1 모델의 핵심인 ’생각하는 과정(Chain-of-Thought)’을 오픈 모델 최초로 내재화하였다. 사용자는 low, medium, high로 추론 강도를 조절할 수 있으며, 모델이 내뱉는 ’사고 과정(Reasoning Path)’ 전체를 투명하게 확인할 수 있다. 이는 블랙박스였던 기존 o1 모델과 달리, 개발자가 AI의 논리적 오류를 디버깅하고 에이전트의 의사결정 과정을 완벽하게 통제할 수 있게 해준다.

(4) Agentic Tool Use - “자율 에이전트 최적화” 185)

단순 텍스트 생성을 넘어, 함수 호출(Function Calling), 웹 브라우징, Python 코드 실행 등 도구 사용 능력이 학습 단계에서 강화되었다. 구조화된 출력(Structured Outputs)을 기본 지원하여, LangChain이나 AutoGPT와 같은 프레임워크에서 복잡한 워크플로우를 수행하는 자율 에이전트(Agent)의 두뇌로 즉시 활용 가능하다.

<표 17> OpenAI gpt-oss 개요

개발자	OpenAI	홈페이지	https://www.eleuther.ai/
최초공개 시기	2021.06 (GPT-J), 2022.02 (GPT-NeoX)	라이선스	Apache License 2.0

183) Kaitchup, 2025, OpenAI GPT-OSS: Native 4-Bit MoE Models, https://kaitchup.substack.com/p/openai-gpt-oss-native-4-bit-moe-models

184) SiliconFlow, 2025, OpenAI's gpt-oss Now Live: Designed for Agentic Workflows, https://www.siliconflow.com/blog/openai-s-gpt-oss-now-live-on-siliconflow-designed-for-agentic-workflows-advanced-reasoning-and-coding

185) GPT-OSS Documentation, 2025, Getting Started Guide: Function Calling, https://www.gpt-oss.io/documentation

<table>
<tr><td>주요 특징</td><td colspan="5">Reasoning(추론) 특화 오픈 모델 / MoE 아키텍처 및 Native MXFP4 양자화 / o3/o4-mini 수준의 코딩 및 수학 성능</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/openai</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>117</td><td>3</td><td>38</td><td colspan="2">13</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>gpt-oss-120b</td><td>8,500,000+</td><td>15,000+</td><td>N/A</td><td>600+</td></tr>
<tr><td rowspan="3">주요 활용 사례</td><td colspan="2">On-Premise Reasoning - 내부망 추론 시스템. 금융/보안 등 데이터 유출이 불가능한 기업들이 o1급 추론 모델을 자사 서버(On-Premise)에 구축. API 비용 없이 고성능 추론 활용.</td><td colspan="3">https://northflank.com/blog/self-host-openai-gpt-oss-120b-open-source-chatgpt

https://thinkaicorp.com/openais-gpt-oss-models-private-cost-effective-reasoning-for-ai-agents/</td></tr>
<tr><td colspan="2">Local Coding Assistant - 로컬 코딩 도우미. 개발자들이 GitHub Copilot 대신 로컬 PC(H100/RTX 4090)에서 독립적으로 구동하는 코딩 AI 구축. 보안 우려 없는 코드 자동 완성 및 리팩토링 수행.</td><td colspan="3">https://www.virtualizationhowto.com/2025/05/best-self-hosted-github-copilot-ai-coding-alternatives/</td></tr>
<tr><td colspan="2">Agentic Workflow - 자율 에이전트 두뇌. 복잡한 도구 호출(Tool Calling)이 필요한 LangChain/AutoGPT 기반 에이전트의 핵심 모델로 채택. 정확한 논리 추론으로 에이전트 성공률 향상.</td><td colspan="3">https://www.langflow.org/blog/openai-gpt-oss-and-gpt-5-on-langflow</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스
가중치 공개 : O (완전 공개) (Hugging Face에서 즉시 다운로드 가능)
코드 공개 : O (완전 공개) (추론/평가/파인튜닝 코드 100% 공개)
데이터 공개 : X (미공개) (학습 데이터셋 구성 비공개 - 경쟁 우위 보호)
라이선스 : Apache 2.0 (수정, 배포, 상업적 판매 모두 자유)
상업적 제한 : 없음 (별도의 사용자 수 제한이나 매출 제한 없음)
OSI 기준 : △ (가중치/코드는 충족하나, 학습 데이터 미공개로 'Open Weight' 분류)</td></tr>
</table>

(자체 작성)

7. 바이두 ERINE

ERNIE 4.5 (Wenxin 4.5)는 중국 최대 검색 엔진 기업인 Baidu(바이두)가 2025년 6월 30일, 기존 폐쇄형 전략을 수정하여 전격 공개한 오픈소스 AI 모델 시리즈[186]

186) Baidu ERNIE Blog, 2025, ERNIE 4.5 Model Family: Open Sourcing MoE Models, https://ernie.baidu.com/blog/posts/ernie4.5/

다. DeepSeek의 성공과 Qwen의 확산 등 중국 내 치열한 오픈소스 경쟁에 대응하기 위해 출시되었으며, 기존 'ERNIE Bot'의 강력한 성능을 경량화된 MoE 아키텍처로 구현하여 개발자 생태계(PaddlePaddle)로의 확장을 꾀하고 있다.

핵심 모델인 ERNIE 4.5 (47B Active)와 경량 모델 ERNIE X1은 Apache 2.0 라이선스로 공개되어 상업적 이용이 가능하다. 특히 바이두의 자체 딥러닝 프레임워크인 PaddlePaddle에 최적화되어 있으며, 2025년 9월 출시된 ERNIE 5.0(Native Omni-modal)으로의 기술적 연결 고리 역할을 수행한다. 이는 "중국 최초의 AI 기업"인 바이두가 오픈 생태계로 회귀함을 상징하는 전략적 모델[187]로 평가받는다.

<표 18> 어니 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
ERNIE 4.5 (2025.06)	424B (47B Active)	- Sparse MoE (고성능/저비용) - PaddlePaddle 최적화 - RAG(검색증강) 특화 학습	128K	- 엔터프라이즈 지식 관리 (KMS) - 중국어 법률/금융 문서 분석 - 대규모 서비스 백엔드
ERNIE X1 (2025.06)	26B (3B Active)	- On-Device MoE (경량화) - 모바일 NPU 가속 지원 - 실시간 음성/텍스트 처리	32K	- 스마트 디바이스 (Xiaodu) - 자율주행 차량 음성 비서 - 엣지 컴퓨팅 환경

(자체 작성)

(1) MoE 아키텍처 (Sparse Mixture-of-Experts) - "효율성의 극대화" [188]

ERNIE 4.5는 총 파라미터 424B 중 토큰당 47B만 활성화하는 Sparse MoE 구조를 채택하였다. 이는 경쟁 모델인 DeepSeek-V3(37B Active)와 유사한 전략으로, 거대 모델의 지능을 유지하면서도 추론 비용을 획기적으로 낮춰 엔터프라이즈 도

187) CNBC, 2025, China's biggest public AI drop since DeepSeek, Baidu's Ernie, is about to hit the market, https://www.cnbc.com/2025/06/29/china-biggest-ai-drop-since-deepseek-baidus-ernie-to-hit-market.html

188) PR Newswire, 2025, Baidu Unveils ERNIE 5.0 and AI Applications at Baidu World 2025, https://www.prnewswire.com/news-releases/baidu-unveils-ernie-5-0-and-a-series-of-ai-applications-at-baidu-world-2025--ramps-up-global-push-302302856.html

입 장벽을 제거하였다. 특히 바이두 클라우드(Qianfan) 상에서 최적화된 추론 속도를 제공한다.

(2) PaddlePaddle 최적화 및 생태계 통합 - “중국형 AI 표준” 189)

대부분의 글로벌 모델이 PyTorch 기반인 것과 달리, ERNIE는 중국의 자체 딥러닝 프레임워크인 PaddlePaddle 기반으로 학습 및 배포되었다. 이는 미국의 기술 제재(GPU 등)에 대비한 중국의 ’기술 자립(Sovereign AI)’ 전략과 맞닿아 있으며, 중국 내 관공서 및 국영 기업(SOE) 시장에서 독점적인 지위를 확보하는 기반이 되고 있다.

(3) 검색 증강 생성 (RAG) 및 에이전트 연동 - “검색의 바이두” 190)

검색 엔진 기업의 강점을 살려, ERNIE 4.5는 외부 지식(Web Search)을 실시간으로 가져와 답변을 보강하는 RAG(Retrieval-Augmented Generation) 능력에 특화되어 있다. 또한, 바이두의 AI 에이전트 서비스인 Wenxiaoyan(文小言)과 긴밀하게 연동되어, 복잡한 사용자 명령을 수행하고 도구를 호출하는 에이전트 성능이 강화되었다.

<표 19> 바이두 어니(ERNIE) 개요

<table>
<tr><td>개발자</td><td colspan="2">Baidu (Baidu AI Cloud)</td><td>홈페이지</td><td colspan="2">ps://ernie.baidu.com</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2025.06.30 (Open Source 전환)</td><td>라이선스</td><td colspan="2">Apache 2.0</td></tr>
<tr><td>주요 특징</td><td colspan="5">MoE (Mixture-of-Experts) 아키텍처 / PaddlePaddle 프레임워크 최적화 / 중국어 NLP 및 검색 증강(RAG) 특화</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/PaddlePaddle</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 멤버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋 수</td></tr>
<tr><td>112</td><td>2</td><td>30</td><td colspan="2">0</td></tr>
<tr><td>핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
</table>

189) AIbase News, 2025, Baidu Open Sources the WENXIN Large Model 4.5 Series, https://news.aibase.com/news/19339

190) AInvest, 2025, Baidu's AI Ecosystem and ERNIE X1.1: A Strategic Catalyst for Long-Term Growth, https://www.ainvest.com/news/baidu-ai-ecosystem-ernie-x1-1-strategic-catalyst-long-term-growth-2509/

	ERNIE-4.5-8B	2,100,000+	4,200	N/A	120+
주요 활용 사례	Samsung China (삼성전자 중국) - 갤럭시 AI 검색.중국 출시 갤럭시 S25 시리즈의 'Circle to Search' 및 AI 비서 기능에 ERNIE 4.5 모델 탑재. (Google Gemini 대안으로 채택)		https://www.cnbc.com/2024/01/26/samsung-taps-baidu-ernie-bot-for-galaxy-s24-ai-features-in-china.html		
	레노버 - AI PC 및 모토로라 스마트폰 내장 AI 어시스턴트에 ERNIE 경량 모델(X1)을 기본 엔진으로 탑재하여 온디바이스 AI 구현		https://research.contrary.com/company/mistral-ai		
	Geely (지리자동차) - 스마트 콕핏(Smart Cockpit). 지리자동차의 프리미엄 전기차 라인업(Zeekr 등)의 차량용 음성 인식 및 인포테인먼트 시스템에 ERNIE 기반 대화형 AI 적용		https://global.geely.com/media-center/news/baidu-collaboration		
	Honor (아너) - 매직OS (MagicOS) 통합. 아너 스마트폰의 운영체제에 ERNIE의 의미 검색 및 요약 기능을 통합하여 사용자 의도를 파악하는 'IWI(Intent-based UI)' 구현		https://www.honor.com/global/news/honor-magic-os-ai		
기술 데이터 공개 범위 및 주요 라이선스 가중치 공개 : O (완전 공개) (Hugging Face 및 PaddlePaddle에서 다운로드) 코드 공개 : O (완전 공개) (PaddleNLP 라이브러리를 통해 전체 공개) 데이터 공개 : X (미공개) (학습 데이터셋은 비공개 정책 유지) 라이선스 : Apache 2.0 (상업적 이용 및 수정 자유) 상업적 제한 : 없음 (별도의 매출 제한 없음) OSI 기준 : △ (가중치/코드 충족, 데이터 미공개로 'Open Weight' 분류)					

(자체 작성)

8. 업스테이지 솔라

Solar (솔라)는 "작지만 강력한(Small but Mighty)" 모델을 지향하는 국내 대표 AI 기업 업스테이지의 플래그십 LLM 시리즈다.

2023년 12월, 10.7B 모델로 세계 최초의 DUS(Depth Up-Scaling) 기술을 선보이며 허깅페이스 리더보드 1위[191]를 차지해 글로벌 주목을 받았고, 2025년 7월에는 31B 크기의 Solar Pro 2를 정식 출시[192]하며 '글로벌 프런티어급(Global Frontier)'

191) Upstage Blog, 2023, Upstage's Solar 10.7B Emerges as World's Top Pre-trained LLM, https://Upstage.ai/news/solar-10-7b-emerges-as-worlds-top-pre-trained-llm

성능을 달성하였다. 특히 자체 평가에서 Intelligence Index 58점을 기록하며 GPT-4o(41점) 및 GPT-4.1(53점)을 능가[193]하였다.

핵심 경쟁력은 “압도적인 가성비와 효율성“이다. 수천억 파라미터(100B~1.7T)를 가진 경쟁 모델들이 거대한 GPU 클러스터를 요구하는 것과 달리, Solar Pro 2는 단일 GPU(NVIDIA A100/H100 1장)만으로도 구동 가능하며, 최적화 시 RTX 3090/4090 같은 소비자용 GPU에서도 실행할 수 있도록 설계[194]되었다. 이러한 특징 덕분에 민감한 데이터를 외부로 내보내지 않고 내부 서버에서 처리하려는 글로벌 기업들의 온프레미스(On-Premise) AI 표준[195]으로 각광받고 있다.

<표 20> 업스테이지 솔라 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
Solar Mini(2023.12)	10.7B	-DUS (Depth Up-Scaling)최초 적용- Llama 2 / Mistral 7B 기반 확장- 동급(13B 이하) 최고 성능	4K / 32K	[가성비/범용]챗봇, 문서 요약, RAG 기본 모델
Solar Pro Preview(2024.09)	22B	-중형 모델의 표준- 싱글 GPU 최적화 (22B의 마법)- GPT-4 대비 2.5배 빠른 속도	32K	[기업용/구축형]사내 지식검색, 금융 약관 분석
Solar Pro 2(2025.07)	31B	-프런티어급 성능 (GPT-4o 준수)-Reasoning(추론) 강화- 한국어/영어 Dual-Language 최적화	32K / 128K	[전문가/고성능]복잡한 추론, 코딩, 법률/의료 자문
Solar Pro 2 (Reasoning)(2025.09)	31B	-System 2 Thinking (사고력 강화)- CoT(Chain of Thought) 내재화- 복잡한 논리/수학 문제 해결 특화	128K	[심층 추론]수학 증명, 전략 기획, 데이터 분석

(자체 작성)

192) Upstage Console, 2025, Solar Pro 2 Released on 2025-07-10, https://console.Upstage.ai/docs/models/solar-pro-2
193) KoreaTechDesk, 2025, Upstage's Solar Pro 2 Beats GPT-4.1 in Global AI Rankings, https://koreatechdesk.com/Upstage-solar-pro-2-korean-global-ai-frontier-model
194) Upstage News, 2025, Solar Pro 2 Breaks Into Global Frontier AI, https://Upstage.ai/news/solar-pro-2-frontier
195) Upstage Pricing, 2025, On-premises Deployment for Enterprise, https://Upstage.ai/pricing/on-premises

(1) DUS (Depth Up-Scaling) - “구조적 효율화의 혁명” [196]

이미 학습된 소형 모델(예: Llama 2 7B, Mistral 7B)의 레이어를 복제하고 재조립(Slicing & Stacking)한 뒤, 추가 학습(Continued Pre-training)을 수행하여 모델의 깊이(Depth)를 늘리는 기술이다. 처음부터 다시 학습하는 비용을 들이지 않고도 모델 용량을 효과적으로 키워 성능을 높인다.

이 방식은 MoE(전문가 혼합)의 대안[197]으로 떠올랐다. 복잡한 MoE 구조 없이도 모델 성능을 2~3배 끌어올릴 수 있음을 증명하여, 중소기업이나 연구소에서도 저비용으로 고성능 LLM을 개발할 수 있는 길을 열었다.

(2) Single-GPU Optimization - “인프라 비용 0원 도전” [198]

Solar Pro 2(31B)는 양자화(Quantization) 기술과 결합 시, 소비자용 그래픽카드인 RTX 4090(24GB VRAM) 한 장에서 추론이 가능하도록 최적화[199]되었다. “AI 도입의 진입장벽“을 허물었다는 평가를 받으며, 고가의 클라우드나 H100 서버 없이도 로컬 PC에서 GPT-4급 지능을 돌릴 수 있게 되어, 데이터 보안이 생명인 금융/의료 분야의 온프레미스 구축 1순위 모델이 되었다.

(3) Purpose-Built Reasoning - “작은 거인의 사고력” [200]

2025년형 Solar Pro 2는 단순 답변 생성을 넘어, DeepSeek-R1과 같은 추론(Reasoning) 능력이 대폭 강화되었다. 특히 한국어/영어 이중언어 환경에서의 복합 추론(예: “한국 법률에 따른 미국 계약서 검토“) 능력은 Llama 3.1 70B를 능가[201]한다.

한국형 전문직 AI(법률, 세무, 특허) 시장에 적용되기 시작했으며, 덩치만 큰 모

196) Hugging Face (업스테이지), 2024, SOLAR-10.7B: Depth Up-Scaling Methodology, https://huggingface.co/업스테이지/SOLAR-10.7B-v1.0
197) Synced Review, 2023, Breaking LLMs' Limits: 업스테이지 AI's SOLAR 10.7B Shines Bright with Simple Scaling Magic, https://syncedreview.com/2023/12/31/breaking-llms-limits-업스테이지-ais-solar-10-7b-shines-bright-with-simple-scaling-magic/
198) 업스테이지 Blog, 2025, Solar Pro 2 Preview: Small. Powerful. Now with reasoning, https://업스테이지.ai/blog/en/solar-pro-2-preview-small-powerful-now-with-reasoning
199) Database Mart, 2025, RTX 4090 LLM Inference Benchmark, https://www.databasemart.com/blog/vllm-gpu-benchmark-dual-rtx4090
200) Upstage Launch News, 2025, Solar Pro 2: Fluent. Reasoning. Frontier., https://Upstage.ai/blog/ko/solar-pro-2-launch
201) Asian Intelligence, 2025, Solar Pro 2: South Korea's Frontier LLM, https://asianintelligence.ai/reports/4/solar-pro-2-south-koreas-frontier-llm

델보다 한국 실정에 맞는 정교한 추론이 가능한 솔라가 실무에 더 적합하다는 평가를 받는다.

<표 21> 업스테이즈 솔라 개요

<table>
<tr><td>개발자</td><td colspan="2">업스테이지 (업스테이지)</td><td>홈페이지</td><td colspan="2">https://www.업스테이지.ai/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2023.12 (Solar 10.7B) / 2025.07 (Solar Pro 2)</td><td>라이선스</td><td colspan="2">Apache 2.0 (Mini, Pro Preview) / Solar Commercial (Pro 2 정식)</td></tr>
<tr><td>주요 특징</td><td colspan="5">DUS 기술로 구현한 압도적 가성비 / 단일 GPU(24GB~) 구동 최적화 / 한국어-영어 이중언어 최강</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/zai-org</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 맴버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋</td></tr>
<tr><td>91</td><td>3</td><td>21</td><td colspan="2">5</td></tr>
<tr><td rowspan="3">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Solar-10.7B-Instruct</td><td>3,500,000+</td><td>5,200+</td><td>N/A</td><td>180+</td></tr>
<tr><td>Solar-Pro-Preview</td><td>850,000+</td><td>1,200+</td><td>N/A</td><td>80+</td></tr>
<tr><td rowspan="4">주요 활용 사례</td><td colspan="3">삼성생명 - 보험금 청구 심사. Solar LLM을 도입하여 복잡한 진단서와 보험 약관을 대조 분석, 지급 심사 자동화율 60% 이상 달성</td><td colspan="2">https://www.업스테이지.ai/blog/en/업스테이지-named-to-cb-insights-ai-100-2025</td></tr>
<tr><td colspan="3">QANDA (콴다) - 수학 문제 풀이. 매스프레소(Mathpresso)와 협력하여 Solar 기반의 'MathGPT'를 개발, 수식 인식 및 풀이 과정 설명 능력 고도화</td><td colspan="2">https://www.업스테이지.ai/news/mathgpt</td></tr>
<tr><td colspan="3">AWS Bedrock - 글로벌 MaaS 공급. 아마존의 AI 플랫폼 Bedrock에 Solar 모델 입점. 전 세계 AWS 고객이 클릭 한 번으로 Solar API 사용 가능</td><td colspan="2">https://www.업스테이지.ai/news/solar-pro-aws</td></tr>
<tr><td colspan="3">ConnectWave (다나와) - 이커머스 상품 추천. 14억 개 상품 데이터를 Solar로 분석하여, 자연어 검색("가벼운데 배터리 오래가는 노트북")에 맞는 최적 상품 추천.</td><td colspan="2">https://www.etnews.com/20240206000282</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스(Solar Mini / Solar Pro Preview)
* Solar Pro 2 (정식) : 클로즈드 모델
가중치 공개 : O (완전 공개) (HuggingFace)
코드 공개 : O (공개) (파인튜닝 예제 및 추론 코드)
데이터 공개 : X (미공개) (학습 데이터 레시피 비공개)
라이선스 : Apache 2.0 (상업적 이용 완전 자유)
상업적 제한 : 없음 (등록 절차 없이 누구나 사용 가능)
OSI 기준 : O (충족)</td></tr>
</table>

(자체 작성)

9. 네이버 HyperCLOVA X

HyperCLOVA X는 네이버가 자체 데이터센터 '각 세종'의 슈퍼컴퓨팅 인프라를 기반으로 개발한 초대규모 AI다.

GPT-4보다 한국어 데이터를 6,500배 더 많이 학습하여 한국의 법률, 규제, 문화적 맥락을 가장 완벽하게 이해한다고 평가받는 소버린 AI의 대표 모델[202]이다.

2025년의 가장 큰 변화는 "개방(Open)"과 "추론(Reasoning)"이다. 그동안 API로만 제공되던 모델을 3B/1.5B/0.5B 경량 버전(SEED)으로 오픈소스화[203]하여 스타트업과 연구소에 무료로 풀었고, 복잡한 논리 문제를 해결하는 System 2 추론 모델(THINK)를 출시하여 단순 챗봇을 넘어 'AI 에이전트'[204]로 진화하였다.

<표 22> 네이버 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
HyperCLOVA X(2023.08)	미공개(초대형)	-한국어 최적화 토크나이저- 한국어/영어 이중언어 학습- 슈퍼컴퓨터 기반 대규모 인프라	32K	[플래그십/범용]네이버 검색(Cue:), 기업용 API, 대화형 에이전트
HCX-DASH(2024.04)	미공개(경량형)	-고효율 경량화 모델- X 모델 대비 1/5 비용으로 90% 성능- 빠른 응답 속도	32K	[가성비/실무]뉴스 요약, 단순 번역, 쇼핑 추천
X-SEED(2025.04)	0.5B / 1.5B / 3B	-최초의 상업용 오픈소스- 온디바이스 최적화- Qwen 2.5 대비 적은 자원으로 고성능	4K	[온디바이스/엣지]모바일 앱, IoT 기기, 로컬 챗봇
X-THINK(2025.06)	미공개(14B 추정)	-추론(Reasoning) 특화- CoT 강화학습- DeepSeek-R1 대응 한국어 모델	32K	[심층 추론]법률 검토, 수학 문제 풀이, 복잡한 계획 수립

(자체 작성)

202) 네이버 CLOVA, 2025, HyperCLOVA X Official Overview, https://clova.ai/en/hyperclova
203) 네이버 Tech Blog, 2025, Introducing HyperCLOVA X SEED: A Commercial Open-Source AI, https://clova.ai/en/tech-blog/sowing-the-seeds-in-the-ai-ecosystem-introducing-hyperclova-x-seed-a-commercial-open-source-ai
204) 네이버 Corporation Press Release, 2025, 네이버 Unveils "HyperCLOVA X THINK," a Reasoning Model with Enhanced AI Agent Capabilities, https://www.네이버corp.com/en/media/pressReleasesDetail?seq=33066

(1) HyperCLOVA X THINK (Reasoning Model) - “한국형 R1의 탄생” [205]

2025년 6월 30일 공개된 추론 특화 모델이다. 사용자의 질문에 바로 답하지 않고, 내부적으로 생각하는 과정(Chain of Thought)을 거쳐 답을 도출한다. DeepSeek-R1과 유사하게 수학, 코딩, 논리 퀴즈에서 압도적인 성능을 보이며, 특히 한국어/영어 이중언어 데이터 6조 토큰을 학습하여 한국어 추론 벤치마크(KoBALT-700 등)에서 최고 수준[206]의 점수를 기록하였다.

기존 LLM이 한국어 말장난이나 복잡한 법률 해석에서 실수를 하였다면, THINK 모델은 “이 법 조항은 A판례와 충돌할 수 있다“는 식의 심층 분석이 가능해져, 국내 로펌의 AI 법률 Q&A 서비스 등 리걸테크(LegalTech)와 금융 분야에 실제 [207]적용되고 있다.

(2) HyperCLOVA X SEED (Open Source) - “생태계의 씨앗”

2025년 4월, 네이버가 처음으로 상업적 이용이 가능한 오픈소스(Apache 2.0 수준)로 공개한 경량 모델군(3B, 1.5B, 0.5B)[208]이다. 이 모델은 “소버린 AI 생태계”의 확장을 목적으로 공개되었으며, 스타트업들이 비싼 API 비용 없이 네이버의 고성능 한국어 모델을 자사 서비스(온디바이스 앱, 로봇 등)에 심을 수 있게 되어, “한국어 AI = 하이퍼클로바” 라는 표준 입지를 굳혔다.

(3) Multimodal Native - “보고 듣고 말하는 한국 AI” [209]

텍스트뿐만 아니라 이미지, 음성을 동시에 이해하는 Multimodal Native 모델로 진화하였다. 한국의 수능 생물 문제 그림을 보고 정답을 맞히는 STEM(과학/기술/공학/수학) 분야 평가에서 46.4% 정확도를 기록하며 GPT-4.1(40.3%)을 앞섰다.[210]

---

205) 중앙일보, 2025, 네이버 unveils homegrown AI model HyperClova X Think, https://koreajoongangdaily.joins.com/news/2025-06-30/business/tech/Naver-unveils-homegrown-AI-model-HyperClova-X-Think/2342061

206) arXiv, 2025, HyperCLOVA X THINK Technical Report, https://arxiv.org/abs/2506.22403

207) Naver Cloud Blog, 2025, HyperCLOVA X Use Cases: Legal Q&A, https://clova.ai/hyperclova

208) Naver Tech Blog, 2025, Introducing HyperCLOVA X SEED, a commercial open-source AI, https://clova.ai/en/tech-blog/sowing-the-seeds-in-the-ai-ecosystem-introducing-hyperclova-x-seed-a-commercial-open-source-ai

209) LinkedIn (Dongsoo Lee), 2025, Naver has officially released its open-source AI model, HyperCLOVA X SEED, https://www.linkedin.com/posts/dongsoo-lee-45028017_Naver-has-officially-released-its-open-source-activity-7320813470411563008-s

210) 네이버 Tech Blog, 2025, HyperCLOVA X THINK: From seeds to forest, https://clova.ai/en/tech-blog/hyperclova-x-think-from-seeds-to-forest

이러한 능력은 네이버의 실생활 서비스에 깊숙이 적용되었다. 2025년형 네이버 지도는 거리뷰 사진을 분석해 정확한 위치 정보를 제공하며, 네이버 쇼핑에서는 “이 옷이랑 어울리는 바지 찾아줘“와 같이 이미지를 통한 맥락 검색 및 상품 추천 기능을 구현[211]하였다.

<표 23> 네이버 하이퍼클로버 X 개요

<table>
<tr><td>개발자</td><td colspan="2">네이버클라우드</td><td>홈페이지</td><td colspan="2">https://clova.ai/hyperclova</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2021.05 (HyperCLOVA) / 2023.08 (X) / 2025.04 (SEED)</td><td>라이선스</td><td colspan="2">네이버 License (X, DASH) / 제한적 라이선스 (SEED - 상업적 이용 가능)</td></tr>
<tr><td>주요 특징</td><td colspan="5">한국어 데이터 학습량 세계 1위 / 소버린 AI 생태계 구축 / 온디바이스(SEED) 및 추론(THINK) 라인업 완성</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/네이버-hyperclovax</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 맴버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋</td></tr>
<tr><td>60</td><td>1</td><td>6</td><td colspan="2">-</td></tr>
<tr><td rowspan="3">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>HyperCLOVAX-SEED-3B</td><td>120,000+</td><td>1,800+</td><td>N/A</td><td>50+</td></tr>
<tr><td>HyperCLOVAX-SEED-Series (누적)</td><td>300,000+</td><td>3,500+</td><td>N/A</td><td>120+</td></tr>
<tr><td rowspan="2">주요 활용 사례</td><td colspan="2">쏘카- 사용자의 모호한 질문(“강원도 가서 차박할 건데 넓은 차 추천해줘“)을 이해하고 적합한 차종과 쏘카존을 추천</td><td colspan="3">https://www.네이버cloudcorp.com/네이버_Cloud_251114_EN.pdf</td></tr>
<tr><td colspan="2">한글과컴퓨터 - 한컴오피스에 하이퍼클로바X를 탑재하여 공문서 초안 작성, 맞춤법 교정, 요약 기능을 제공</td><td colspan="3">https://www.koreatimes.co.kr/business/companies/20250423/네이버-to-release-3-ai-models-as-open-source-free-for-commercial-use</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스(HyperCLOVA X SEED)
*HyperCLOVA X / DASH / THINKSMS 클로즈드 모델
가중치 공개 : O (완전 공개) (HuggingFace)
코드 공개 : O (일부 공개) (추론용 코드 제공)
데이터 공개 : X (미공개)
라이선스 : 네이버 Open License (상업적 이용 허용, 재배포 시 명시 필수)
상업적 제한 : 없음 (단, 유해한 용도 금지)
OSI 기준 : △ (부분 충족) (Llama와 유사한 커스텀 오픈 라이선스)</td></tr>
</table>

(자체 작성)

211) Korea Bizwire, 2025, 네이버 Unveils Major AI Upgrades with HyperCLOVA X for Search, Shopping and Maps, http://koreabizwire.com/Naver-unveils-major-ai-upgrades-with-hyperclova-x-for-search-shopping-and-maps-at-dan24-conference/29716

10. LG 엑사원

EXAONE(Expert AI for Everyone)은 LG 그룹이 “전문가 수준의 AI“를 목표로 개발한 초거대 AI 모델이다.

2024년 8월, 7.8B 경량 모델을 한국 최초로 오픈소스(CC-BY-NC-SA)로 공개[212] 하며 기술력을 입증하였다.

가장 큰 특징은 “압도적인 데이터 품질” 이다. 논문, 특허, 코드, 화학식 등 6천만 건의 전문 데이터를 집중 학습[213]하여, 일반 상식뿐만 아니라 화학, 바이오, 기계공학 분야의 전문 지식에서 타 모델을 압도[214]한다.

2025년에는 구글 클라우드와의 파트너십[215]을 통해 전 세계 연구소와 기업이 EXAONE을 클라우드 API 및 마켓플레이스에서 즉시 사용할 수 있게 되었다.

<표 24> 엑사원 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟	권장 하드웨어/배포 환경
EXAONE 3.0 7.8B(2024.08)	7.8B	-오픈소스 공개 (연구용)- 이중언어(한/영) 특화 학습- 8T 토큰 고품질 데이터 학습	8K	[연구/학술] 논문 분석, 특허 검색, 전문 번역	소비자용 GPURTX 3090 / 4090 (24GB)
EXAONE 3.0 Light(2024.08)	미공개(경량)	-실시간 처리 최적화- 응답 속도 극대화- 모바일/임베디드 적용 가능	4K	[온디바이스] 모바일 비서, 가전제품 제어	엣지 디바이스NPU 탑재 기기
EXAONE 3.5(2025.03)	미공개(중대형)	-추론(Reasoning) 능력 강화- 긴 문맥(Long Context) 처리 개선-	32K	[엔터프라이즈] 기업용 챗봇 (ChatEXAONE),	GPU 서버H100 이상

212) LG AI Research GitHub, 2024, EXAONE 3.0 License & Repository, https://github.com/LG-AI-EXAONE/EXAONE-3.0

213) Korea JoongAng Daily, 2024, LG AI Research launches Korea's first open-source AI model, https://koreajoongangdaily.joins.com/news/2024-08-07/business/industry/LG-AI-Research-launches-Koreas-first-opensource-AI-model-Exaone-30/2107568

214) LG AI Research Technical Report, 2024, EXAONE 3.0 7.8B Instruction Tuned Model Performance, https://github.com/LG-AI-EXAONE/EXAONE-3.0/blob/main/README.md

215) Google Cloud Press Release, 2024, LG AI Research Taps Google Cloud to Develop EXAONE 3.0 and ChatEXAONE, https://www.googlecloudpresscorner.com/2024-08-28-LG-AI-Research-Taps-Google-Cloud-to-Develop-EXAONE-3-0-and-ChatEXAONE-AI-Agent

		RAG 성능 고도화		법률 검토	

(자체 작성)

(1) Bilingual & Hybrid Mastery - “언어와 시각을 동시에 완벽하게” 216)

기존 3.0 모델이 한국어와 영어를 1:1 수준으로 완벽하게 구사하여 Llama 3.1을 능가하는 이중언어(Bilingual) 능력을 보여주었다면, 2025년형 EXAONE 4.0은 텍스트와 이미지를 동시에 이해하는 하이브리드(Hybrid) AI로 진화하였다.

영어로 된 최신 논문을 읽고 한국어로 요약하는 것을 넘어, 복잡한 도면이나 화학 구조식 이미지까지 인식하여 분석한다. 실제 벤치마크에서 글로벌 경쟁 모델 대비 전문 문서 독해 및 시각 정보 처리 정확도가 대폭 향상되어, 언어 장벽뿐만 아니라 데이터 형태의 장벽까지 허물었다.

(2) Industrial Specialization - “화학/특허/제조를 이해하는 산업 특화 두뇌” 217)

LG화학, LG디스플레이 등 계열사의 실제 R&D 데이터를 학습한 노하우는 2025년 엑사원 생태계로 확장되었다. 7.8B 모델이 특허 검색 시간을 1/100로 줄인 성과를 바탕으로, 이제는 신약 후보 물질 발굴이나 배터리 불량 검출 같은 고난도 제조/바이오 영역까지 커버한다.

특히 2025년 11월 오픈 베타를 시작한 기업용 AI 에이전트 ChatEXAONE(챗엑사원)은 ISO 보안 인증을 획득하여, 국가 핵심 기술이나 기업 내부의 민감한 데이터를 다루는 보안이 보장된 R&D 파트너로 자리 잡았다.

(3) Efficient Architecture & Data Foundry - “비용은 1/4, 생산성은 1,000배” 218)

이전 대비 추론 비용을 72% 절감하고 로컬 PC(RTX 4090) 구동을 실현한 3.0의 효율성에 더해, 2025년에는 ’엑사원 데이터 파운드리(Data Foundry)’를 통해 효율

216) Chosun Biz, 2025, LG unveils EXAONE ecosystem and hybrid AI at AI Talk Concert 2025, https://biz.chosun.com/en/en-industry/2025/07/22/YIQB4QQ5O5ED5JSFXXR6F6C3LA/

217) LG AI Research Official YouTube, 2025, Meet ChatEXAONE: LG’s Enterprise AI Agent, https://www.youtube.com/watch?v=asQhZI3_eTo

218) Korea Herald, 2025, LG AI Research Showcases 'EXAONE Ecosystem', https://www.koreatimes.co.kr/business/tech-science/20250722/lg-exaone-data-foundry

의 차원을 넓혔다.

단순히 모델만 가벼운 것이 아니라, AI 학습에 필요한 고품질 데이터를 자동으로 생성해주는 기술을 도입하였다. 이를 통해 전문가 60명이 3개월 걸리던 데이터 가공 작업을 단 하루 만에 끝낼 수 있게 되어(생산성 1,000배 향상), 기업들이 AI를 도입하고 고도화하는 데 드는 시간과 비용을 획기적으로 낮췄다.

<표 25> LG 엑사원 3.0 개요

<table>
<tr><td>개발자</td><td colspan="2">LG AI Research (LG AI연구원)</td><td>홈페이지</td><td colspan="2">https://www.lgresearch.ai/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2021.12 (1.0) / 2024.08 (3.0) / 2025.03 (3.5)</td><td>라이선스</td><td colspan="2">CC-BY-NC-SA 4.0</td></tr>
<tr><td>주요 특징</td><td colspan="5">화학/특허 등 전문 데이터 6천만 건 학습 / 한국어-영어 이중언어 성능 최강 / 구글 클라우드 파트너십</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://huggingface.co/LGAI-EXAONE</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 맴버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋</td></tr>
<tr><td>6</td><td>7</td><td>39</td><td colspan="2">5</td></tr>
<tr><td rowspan="2">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>EXAONE-3.0-7.8B-Instruct</td><td>125,000+</td><td>2,100+</td><td>N/A</td><td>60+</td></tr>
<tr><td rowspan="2">주요 활용 사례</td><td colspan="2">LG화학 - 신소재 발굴. 논문과 특허에 있는 화학 구조식을 EXAONE이 분석하여, 신소재 후보 물질을 추천</td><td colspan="3">https://www.chosun.com/english/industry-en/2024/08/07/JDJ7GFFMXBAWXPZW3VRTCU3SCU/</td></tr>
<tr><td colspan="2">Shutterstock - 이미지 캡셔닝. EXAONE의 멀티모달 기능을 활용해 수억 장의 이미지에 자동으로 고품질 설명을 달아 검색 정확도 향상</td><td colspan="3">https://www.googlecloudpresscorner.com/2024-08-28-LG-AI-Research-Taps-Google-Cloud-to-Develop-EXAONE-3-0-and-ChatEXAONE-AI-Agent</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스
가중치 공개 : O (완전 공개) (HuggingFace에서 승인 후 다운로드)
코드 공개 : O (공개) (추론 및 튜닝용 코드 제공)
데이터 공개 : 부분 공개 (KoMT-Bench 등 평가 데이터셋 공개, 학습 데이터 비공개)
라이선스 : CC-BY-NC-SA 4.0 (비상업적 용도 무료 / 상업적 이용 시 LG와 계약 필요)
상업적 제한 : 있음 (연구/개인 목적 외 상업적 사용 불가 - 오픈 웨이트 모델)
OSI 기준 : △ (부분 충족) (Llama와 유사한 커스텀 오픈 라이선스)</td></tr>
</table>

(자체 작성)

11. SK 텔레콤 A.X

A.X (에이닷 엑스)는 SK텔레콤의 AI 전략인 '글로벌 AI 컴퍼니(Global AI Company)' 도약을 위한 핵심 모델이다. 2025년 7월, 최신 버전인 A.X 4.0을 공개하며 글로벌 경쟁력을 입증하였다.

가장 큰 특징은 "검증된 글로벌 SOTA(State-of-the-Art) 오픈소스를 한국형으로 완벽하게 재해석"[219]하였다는 점이다. 자체 개발만을 고집하지 않고, 알리바바의 Qwen 2.5 아키텍처를 과감히 도입[220]한 뒤, SK 텔레콤이 보유한 방대한 고품질 한국어 데이터로 지속적 사전학습(CPT)과 미세조정(SFT)을 수행하였다.

그 결과, A.X 4.0은 GPT-4o보다 한국어 토큰 처리 효율이 33% 더 높은 압도적인 경제성을 달성하였다. 또한, 한국어 성능의 척도인 KMMLU 벤치마크에서 78.3점을 기록[221]하며 GPT-4o(72.5점)를 능가, '외산 베이스 모델 + 국산 고품질 데이터' 조합의 성공 방정식을 증명하였다.

현재 72B(Standard)와 7B(Light) 두 가지 버전으로 제공되며, SK텔레콤의 AI 인프라 슈퍼하이웨이(AI Infrastructure Superhighway) 전략에 따라 B2B 엔터프라이즈 시장과 에이닷(A.) 서비스의 핵심 두뇌로 활용[222]되고 있다.

<표 26> SK텔레콤 A.X 주요 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
A.X 3.1 (2025.07)	34B	-자체 개발(Sovereign AI)- 2.1T 토큰 고효율 학습- 한국어 뉘앙스/문화 완벽 이해	32K	[기술 자립/보안]정부 기관, 금융권 내부망 구축
A.X 4.0	7B	-Qwen 2.5 기반 튜닝- 온디바이스	16K	[온디바이스/엣지]모

219) Hugging Face (SK 텔레콤/A.X-4.0), 2025, A.X 4.0 Model Card & Performance, https://huggingface.co/SK 텔레콤/A.X-4.0

220) SK Telecom Press Release, 2025, SK Telecom Unveils Proprietary Standard Large Language Model A.X 4.0, https://news.SK 텔레콤elecom.com/en/2035

221) Pulse News (Maeil Business Newspaper), 2025, SK telecom unveils Korean-optimized AI model 'A. X 4.0', https://pulse.mk.co.kr/news/english/11358541

222) SK Telecom Newsroom, 2025, SK Telecom Unveils Plans for 'AI Infrastructure Superhighway', https://news.SKtelecom.com/en/1606

Light (2025.07)		최적화- 빠르고 가벼운 상담 요약		바일 에이닷 앱, 상담원 보조 툴
A.X 4.0 Standard (2025.07)	72B	-플래그십 성능 (Qwen 72B 기반)- KMMLU 78.3점 (GPT-4o 상회)- 복잡한 문서 처리 및 추론	128K	[엔터프라이즈]복잡한 민원 처리, 법률/규정 검토
A.X 4.0 VL (2025.07)	미공개 (Vision)	-시각-언어 멀티모달- 문서 이미지 인식(OCR) 통합- 표/차트 해석 능력 강화	32K	[문서 자동화]청구서 인식, 신분증 진위 확인

(자체 작성)

(1) Qwen 2.5-Based Optimization - 과감한 기존 오픈소스 모델 활용[223]

A.X 4.0은 Qwen 2.5를 베이스 모델로 채택하였다. Qwen의 압도적인 코딩/수학 능력은 그대로 유지하면서, SK 텔레콤이 자체 개발한 '한국어 토크나이저(Tokenizer)'를 이식하여 한국어 처리 속도와 효율성을 획기적으로 개선하였다.

이로인해 "개발 기간 단축과 성능 극대화"를 동시에 이뤘다. 바닥부터 만드는 비용을 아끼는 대신, 그 자원을 데이터 품질 향상에 쏟아부어 네이버 HyperCLOVA X와 대등한 한국어 성능을 훨씬 적은 비용으로 구현하였다.

(2) Token Efficiency Revolution - "운영 비용 33% 절감" [224]

설명: 기존 Qwen 모델이나 GPT-4o가 한국어 한 문장을 처리할 때 20개의 토큰을 쓴다면, A.X 4.0은 자체 토크나이저 덕분에 13~14개만 사용[225]한다.

API 호출 비용이 곧 경쟁력인 B2B 시장에서, 똑같은 성능이면 더 싸고 빠른 A.X 4.0이 스타트업과 중소기업의 최우선 고려 대상이 되고 있다.

<표 27> SK텔레콤 A.X 개요

개발자	SK텔레콤	홈페이지	https://www.SK

223) SK Telecom Official Press Release, 2025, SK Telecom Unveils Proprietary Standard Large Language Model A.X 4.0, https://news.SKtelecom.com/213534

224) HelloT (IT News), 2025, SK 텔레콤 unveils A.X 4.0: Higher token efficiency than GPT-4o, https://www.hellot.net/news/article.html?no=102887

225) DealSite, 2025, SK Telecom's A.X 4.0 beats GPT-4o in Korean efficiency, https://dealsite.co.kr/articles/143106/068020

<table>
<tr><td></td><td colspan="2"></td><td>telecom.com/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2023.09 (A.X 선언) / 2025.07 (4.0 오픈소스 공개)</td><td>라이선스</td><td>Apache 2.0 (A.X 4.0) / SK 텔레콤 Research License (A.X 3.1)</td></tr>
<tr><td>주요 특징</td><td colspan="4">Qwen 2.5 기반의 압도적 가성비 / 한국어 토큰 효율 33% 개선 / 자체 개발(3.1)과 오픈 튜닝(4.0) 병행</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="4">https://huggingface.co/SK 텔레콤</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 맴버</td><td>컬렉션 수</td><td>모델 수</td><td>데이터셋</td></tr>
<tr><td>37</td><td>4</td><td>11</td><td>3</td></tr>
<tr><td rowspan="3">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>A.X-4.0-Light</td><td>32,000+</td><td>450+</td><td>N/A</td><td>25+</td></tr>
<tr><td>A.X-3.1-Standard</td><td>15,000+</td><td>300+</td><td>N/A</td><td>15+</td></tr>
<tr><td rowspan="3">주요 활용 사례</td><td colspan="2">SK Telecom (에이닷) - 개인화 비서. 통화 녹음 요약, 일정 자동 등록, 수면 분석 등 에이닷 앱의 모든 텍스트 처리 기능을 A.X 4.0 Light가 전담</td><td colspan="2">https://news.SK 텔레콤elecom.com/213534</td></tr>
<tr><td colspan="2">SK Hynix - R&D 지식 검색. 반도체 공정 관련 방대한 기술 문서를 A.X 4.0 72B가 학습하여, 연구원들의 질문에 전문적인 답변 제공</td><td colspan="2">https://www.chosun.com/english/industry-en/2025/10/30/JC2QOTOPSRBYHJ7VZF5FXA4LR4/</td></tr>
<tr><td colspan="2">하나금융 - AI 상담원. A.X의 빠르고 정확한 한국어 이해 능력을 활용해 고객 상담 챗봇의 응답 정확도를 높이고 상담원 업무 부하 경감</td><td colspan="2">https://www.telecoms.com/ai/sk-telecom-takes-on-openai-with-updated-llm</td></tr>
</table>

기술 데이터 공개 범위 및 주요 라이선스

구분	A.X 4.0 (Standard / Light)	A.X 3.1
가중치 공개	O (완전 공개)(HuggingFace)	O (완전 공개)
코드 공개	O (공개)(추론 및 튜닝 코드)	O (공개)
데이터 공개	X (비공개)(학습 데이터셋 미공개)	X (비공개)
라이선스	Apache 2.0(상업적 이용 완전 자유)	SK 텔레콤 Research License(연구 목적 한정)
상업적 제한	없음(누구나 무료로 서비스 개발 가능)	있음(별도 계약 필요)
OSI 기준	O (충족)	X (불충족)

(자체 작성)

12. 엔씨AI VARCO

VARCO (Via AI, Realize your Creativity and Originality)는 엔씨소프트(NCSoft)가 개발한 LLM의 이름이다.[226] 초기에는 게임 시나리오 작성과 NPC의 실감 나는 대화 생성을 목적으로 개발되었으나, 2024년 9월 Llama 3.1을 기반으로 한국어 성능을 극한으로 끌어올린 'Llama-VARCO' 시리즈를 오픈소스로 공개하며 범용 LLM 시장에서도 두각[227]을 나타내고 있다.

가장 큰 특징은 "창의적인 글쓰기와 논리적 추론(Reasoning)" 의 조화[228]다. 딱딱한 정보 전달 위주의 타 모델과 달리, 소설 작성이나 캐릭터 페르소나(Persona) 이입, 매끄러운 번역에 강점을 보인다.

특히 100억 개(10B) 이하 파라미터 모델 중 한국어 추론 능력을 평가하는 LogicKor 벤치마크 1위를 기록하며, "작지만 똑똑한 모델"로서의 입지를 굳혔다.

2025년 7월에는 텍스트뿐만 아니라 이미지와 영상까지 이해하는 멀티모달 모델[229]인 VARCO-VISION 2.0을 연이어 오픈소스로 공개[230]하며, 게임 개발뿐만 아니라 다양한 산업군에서 활용 가능한 AI 생태계를 확장하고 있다.

<표 28> NC AI의 주요 VARCO 모델

모델명 (공개 시기)	파라미터	아키텍처 및 핵심 기술 혁신	컨텍스트	주요 용도 및 타겟
Llama-VARCO -8B(2024.09)	8B	-Llama 3.1 기반 한국어 튜닝 - LogicKor(한국어 추론) 동급 1위- DPO(선호도 최적화) 적용	8K	[범용/개발자]한국어 챗봇, RAG, 문서 요약
VARCO LLM 2.0(2024.05)	14B / 52B	-자체 개발(4개 국어)- 한/영/중/일 다국어 특화- 창의적	4K	[콘텐츠/게임]게임 NPC 대화, 글로벌

226) NCSoft Official Newsroom, 2024, NCSOFT Unveils 'Llama-VARCO LLM', https://about.ncsoft.com/en/news/article/news_update_240926
227) Hugging Face (NCSoft), 2025, Llama-VARCO-8B-Instruct Model Card, https://huggingface.co/NCSOFT/Llama-VARCO-8B-Instruct
228) The Elec, 2024, NCSoft accelerates AI solution business with Llama-VARCO, https://www.thelec.kr/news/articleView.html?idxno=30320
229) Hugging Face (NCSoft), 2025, VARCO-VISION-2.0-14B Release Note, https://huggingface.co/NCSOFT/VARCO-VISION-2.0-14B
230) NC AI Blog, 2025, VARCO-VISION 2.0 Open Source Release, https://marcus-story.tistory.com/230

		시나리오 작성 능력		CS, 번역
VARCO Vision 2.0(2025.09)	1.7B	-초경량 멀티모달- 이미지-텍스트 통합 이해- 모바일 구동 가능한 사이즈	32K	[온디바이스 비전] 이미지 검색, 게임 QA 자동화

(자체 작성)

(1) Llama-VARCO-8B-Instruct - “한국어 튜닝의 정석” 231)

메타의 Llama 3.1 8B 모델에 NCSoft가 자체 구축한 고품질 한국어 데이터셋을 추가 학습(Continuous Pre-training)시켜 탄생한 모델이다.

LogicKor 벤치마크에서 10B 이하 파라미터 모델 중 1위를 차지할 만큼 압도적인 성능을 자랑한다. 경량 모델임에도 복잡한 한국어 추론과 글쓰기에 능통하여, 국내 개발자들 사이에서 “가장 실용적인 한국어 베이스 모델“로 평가받는다.

(2) VARCO LLM 2.0 - “다국어 마스터” 232)

한국어와 영어뿐만 아니라 중국어, 일본어까지 4개 국어를 원어민 수준으로 구사하는 모델이다. 글로벌 동시 출시가 필수인 게임 산업의 특성을 반영하여 개발되었다. 단순 번역을 넘어 게임 내 고유 명사와 뉘앙스를 정확히 살려내며, 실시간 채팅 번역이나 다국어 고객 응대(CS) 시스템에서 구글 번역기보다 우수한 성능을 발휘한다.

(3) VARCO Judge - “AI를 평가하는 AI” 233)

다른 LLM이 내놓은 답변의 품질을 채점하는 ’평가 전용 모델(Judge LLM)’**이다. 국내 최초로 출시된 이 모델은 기업들이 자사 챗봇 성능을 측정할 때, 비싼 GPT-4 API 대신 무료로 사용할 수 있는 경제적인 대안을 제시하였다. 세계적 권위의 학회인 EMNLP 2024에 논문이 채택되며 기술적 신뢰성도 검증받았다.

231) NCSoft Official Newsroom, 2024, NCSOFT Unveils 'Llama-VARCO LLM', https://about.ncsoft.com/en/news/article/news_update_240926
232) NC Research Tech Blog, 2024, VARCO LLM 2.0 Details, https://ncsoft.github.io/ncresearch/varco-llm-details/
233) Korea Economic Daily, 2024, NCSOFT launches evaluation model to verify performance, https://www.kedglobal.com/artificial-intelligence/newsView/ked202409230015

(4) Manufacturing Digital Twin - “게임 기술을 공장으로” [234]

2025년 2월 분사한 NC AI는 게임 그래픽 제작 기술인 **’VARCO 3D’**를 제조업의 디지털 트윈(Digital Twin) 분야로 확장하였다.

텍스트나 이미지를 3D 에셋으로 빠르게 변환하는 기술을 활용하여, 항공기 부품 공장의 설비 시뮬레이션이나 로봇 가상 시운전 환경을 구축하는 데 적용되고 있다. 게임을 위해 만든 AI가 이제는 스마트 팩토리의 생산성을 높이는 핵심 인프라로 진화한 것이다.

<표 29> NC AI(구, NC소프트) VARCO 개요

<table>
<tr><td>개발자</td><td colspan="2">NCSOFT (엔씨소프트) - NC AI</td><td>홈페이지</td><td colspan="2">https://ncsoft.github.io/ncresearch/</td></tr>
<tr><td>최초공개 시기</td><td colspan="2">2023.08 (VARCO 1.0) / 2024.09 (Llama-VARCO)</td><td>라이선스</td><td colspan="2">Llama Community License (Llama-VARCO) / 자체 라이선스 (VARCO 2.0)</td></tr>
<tr><td>주요 특징</td><td colspan="5">한국어 추론(LogicKor) 1위 / 4개 국어 지원 / 게임 및 콘텐츠 창작에 최적화된 톤앤매너(TTS)</td></tr>
<tr><td>허깅페이스 주소</td><td colspan="5">https://ncsoft.github.io/ncresearch/</td></tr>
<tr><td rowspan="2">허깅페이스 현황</td><td>팀 맴버</td><td>컬렉션 수</td><td>모델 수</td><td colspan="2">데이터셋</td></tr>
<tr><td>29</td><td>4</td><td>9</td><td colspan="2">6</td></tr>
<tr><td rowspan="3">핵심 모델 개발 지표</td><td>모델명</td><td>다운로드 수</td><td>Likes 수</td><td>팔로워 수</td><td>community 수</td></tr>
<tr><td>Llama-VARCO-8B-Instruct</td><td>42,000+</td><td>550+</td><td>N/A</td><td>20+</td></tr>
<tr><td>VARCO-LLM-2.0-7B</td><td>15,000+</td><td>250+</td><td>N/A</td><td>10+</td></tr>
<tr><td rowspan="3">주요 활용 사례</td><td colspan="2">NCSOFT (리니지/아이온) - 게임 콘텐츠 생성. NPC 대사 자동 생성, 퀘스트 시나리오 초안 작성, 실시간 다국어 채팅 번역</td><td colspan="3">https://about.ncsoft.com/en/news/article/news-update-230816</td></tr>
<tr><td colspan="2">VARCO Studio - 일반 유저들이 소설, 시나리오, 웹툰 콘티를 쉽게 구성하도록 도와주는 AI 창작 플랫폼.</td><td colspan="3">https://koreatechdesk.com/ncsoft-unveils-varco-llm-a-versatile-ai-language-model-for-creative-and-original-applications</td></tr>
<tr><td colspan="2">POSTECH - 로봇 자연어 제어용으로 사람의 모호한 명령을 로봇의 구체적 행동 코드로 번역하는 LLM 에이전트 공동 개발</td><td colspan="3">https://www.thisisgame.com/articles/176188</td></tr>
<tr><td colspan="6">기술 데이터 공개 범위 및 주요 라이선스
구분 | Llama-VARCO-8B | VARCO LLM 2.0</td></tr>
</table>

234) Maeil Business Newspaper, 2025, NC AI announced VARCO 3D as core infrastructure for digital twin, https://www.mk.co.kr/en/it/11481131

가중치 공개	O (완전 공개) (HuggingFace)	O (완전 공개) (AWS, Github 등)
코드 공개	O (공개)	X (일부 공개)
데이터 공개	X (비공개)	X (비공개)
라이선스	Llama 3.1 Community License (상업적 이용 가능)	VARCO License (연구용 무료, 상업용 유료)
상업적 제한	없음 (Llama 정책 따름)	있음 (계약 필요)
OSI 기준	△ (부분 충족)	X (불충족)

제5절 요약 및 시사점

1. 요약

본 장은 오픈소스AI 생태계를 기업과 대표 모델 관점에서 동향을 조사하였다. 주요 조사 항목들은 국내외 기업들의 오픈소스AI 관련 현황으로 관련 사업 현황과 주요 오픈소스 모델 현황들을 통해 개별 기업들의 오픈소스 전략, 관련 사업 및 수익화 현황, 오픈소스 모델의 주요 기술적 특징, 핵심 모델 현황(주요 특징, 허깅페이스 현황, 주요 활용 사례, 공개 항목 현황)들을 조사하여 국내외 오픈소스AI 현황을 제시하고 있다.

대상 기업으로 글로벌 오픈소스AI 생태계를 선도하는 해외 8개 기업(메타, 구글, OpenAI, EleutherAI, 알리바바, 바이두, 딥시크, 미스트랄AI)과 국내 오픈소스 AI를 선도하는 기업 5개(네이버, LG AI Research, SK텔레콤, 업스테이지, 엔씨 AI)들이다. 분석 결과, 오픈소스AI는 단순한 연구 성과 공개를 넘어 플랫폼 주도권, 개발자 생태계 확보 및 확산(클라우드, 온디바이스 등)을 위한 핵심 경쟁 수단으로 기능하고 있다. 특히, 2025년에는 미국 기업 중심의 오픈소스AI 생태계에서 딥시크의 공개 이후 중국 기업들의 오픈소스 모델이 글로벌 영향력을 빠르게 확대해 나가면서 미・중 AI 기술 경쟁이 격화되고 있는 현상이 뚜렷해지고 있다.

글로벌 기업 동향의 주요 특징은 다음과 같다. 첫째, 메타・구글 등 빅테크는 오픈 모델을 통해 자사 플랫폼(클라우드・검색・안드로이드・디바이스) 영향력을 확대하며 기술 영향력을 확대하고 있다. 둘째, 중국(알리바바・딥시크・바이두)은 관대한 라이선스(아파치, MIT 등)와 비용 효율성을 전면에 내세워 빠르게 영향력을 확대하며, 산업 적용 사례를 통해 산업적 활용을 확대해 나가고 있다. 셋째, 유럽(미스트랄AI)은 데이터 주권과 규제 준수 수요를 기반으로 유럽 공공・기업 시장을 공략하며 유럽 지역의 AI 영향력 확보를 추진하고 있다.

글로벌 오픈소스 모델 동향을 보면, 메타 라마는 누적 다운로드 10억 회 이상과 높은 점유율을 기반으로 오픈소스 모델 진영의 사실상 표준에 가까운 영향력을 구축해왔다. 하지만, 제한적 공개라는 오픈 웨이트 중심 공개, 상업적 활용의 일부 제한 등으로 엄격한 의미의 오픈소스AI와는 구분되며 오픈소스 진영의 비판을 받아왔다. 그리고, 라마를 활용하여 자사 SNS 서비스의 AI 기능 강화, 상용 AI 서비스 제공, 라마 개발 도구 제공, 신뢰성 보장 등을 통해 수익화를 추진하고 있

다.

구글은 젬마와 제미나이(Gemini)를 연계하여 외부 생태계 확산은 젬마가 담당하고 자체 서비스와 수익화는 제미나이가 담당하는 이중 전략을 추진하고 있다. 이를 통해 제미나이 서비스 제공, 안드로이드 기반 온디바이스 AI 적용, 기존 구글 서비스와 AI 연계 등을 통해 자체 서비스 혁신과 수익화를 동시에 추진하고 있다. OpenAI는 과거 공개 정책에서 GPT-3 이후 폐쇄형 정책을 추진하였으나 딥시크 공개 이후 오픈소스 모델 영향력이 확대되면서 이를 견제하기 위해 gpt-oss(오픈 웨이트)를 공개하며 생태계 영향력을 유지하기 위해 노력하고 있다.

중국에서는 알리바바 큐웬이 이 다국어 지원, 장문 컨텍스트, MoE 등 기술 고도화하며 우수한 성능을 제공하고 있으며, 아파치 라이선스를 통해 자유로운 활용을 허용하면서 미국 중심의 AI 생태계에 균열을 일으키고 있다. 그리고 알리바바도 미국 기업처럼 자체 플랫폼 서비스에 AI 기능을 추가하거나 내부적으로 AI를 활용하여 생산성을 향상시키고, 허깅페이스를 통해 글로벌 영향력을 확대해 나가고 있다.

바이두도 알리바바처럼 오픈소스 라이선스를 적용한 모델을 공개하고 있으며, 이를 기반으로 자체 AI 서비스, 아폴로 프로젝트와 연계된 자율주행, 바이두 지도 서비스와 연계 등과 같은 자체 서비스 혁신 및 AI 서비스를 제공하고 있다. 딥시크는 고성능-저비용 기술 혁신을 MIT 라이선스 기반 모델 공개로 크게 주목을 받았으며, 알리바바와 바이두와 달리 자체 플랫폼이 없기 때문에 우수한 기술력을 기반으로 비즈니스 기회를 적극 창출해 나가고 있다.

이에 비해 국내 기업들은 상대적으로 글로벌 주목도는 다소 낮지만, 2025년부터 적극적을 모델을 공개하며 생태계 영향력을 키우고 있다. 특히 정부의 독자 AI 파운데이션 모델 개발 사업에 적극 참여하며 이를 연계하여 산업 분야의 AI 적용을 위한 경량형 모델을 중심으로 차별화를 시도하고 있다.

오픈소스 모델 현황을 살펴보면, 모델 공개는 단순히 기술 공개가 아니라 선도 기업 견제, 생태계 영향력 확보, 개방형 기술력 검증을 통한 기술·산업 주도권 확보 경쟁의 수단이다. 과거에는 미국 중심의 오픈 웨이트 전략(제한적 공개)으로 재현성을 일부 제한하며 독점적 지위를 유지하기 위해 노력했지만, 이는 중국 기업들의 오픈소스 라이선스를 채택한 개방형 공개로 미국의 AI 주도권이 위협받고 있는 상황이다.

그리고 모델을 공개한 국내외 기업들은 단순히 기술 공개에 머무르지 않고 우선 자체 플랫폼/서비스/제품의 고도화(AI 기능 추가)를 추진하고 있으며, AI 서비스를 제공하며 수익 및 새로운 데이터 수집을 추진하고, AI 솔루션을 제공하며 새로운 비즈니스 수익을 창출해 나가고 있다.

2. 시사점

1) 오픈소스AI 전략의 실질적 목표 : 기술·산업 영향력 확대

메타, OpenAI 등 글로벌 AI 선도 기업들은 자사 기술을 공개하며 생태계 기여, AI 기술 민주화 라는 선의의 목적을 전면에 내세우고 있다. 하지만, 이들 기업들이 오픈소스AI를 공개하는 실질적 이유를 기술 공개 시점, 공개 방법, 경쟁 기업 대응 등을 통해 추정하면 이러한 선의의 목적만 있지 않음을 유추할 수 있다.

메타의 경우 라마를 전략적으로 공개한 시점은 바로 OpenAI가 chatGPT 서비스를 출시하며 AI 서비스 시장에서 독점적 지위를 확보하려던 시점이었다. 이때 메타는 라마를 오픈 웨이트 형태의 연구 목적용으로 제한적 공개하였다. 그리고 이후 상업적 활용을 허용하면서 월간 7억명 이상은 사용 금지하고 있다. 이러한 제약에도 비싼 chatGPT 비용과 자유로운 연구를 위해 많은 기업과 대학들이 라마를 활용하면서 빠른 기술 혁신을 통한 기술 격차를 해소할 수 있었다. 이와 같은 공개 전략은 GPT의 대안 기술을 제공하며 OpenAI의 독점적 지위를 견제하고 빠른 기술 혁신을 위한 전략적 목적으로 가지고 있음을 쉽게 알 수 있다.

OpenAI는 GPT-3 이전인 GPT-1과 GPT-2를 공개하며 자체 기술의 우수성을 전 세계에 증명하며 AI 생태계에서 주목받는 비영리 기관이었다. 하지만. 2022년 chatGPT를 출시하며 영리 기업으로 전환하고 본격적인 수익화를 위해 후속 모델들을 비공개로 전환하였다. 만약 OpenAI가 chatGPT 서비스 출시 이전에 모델을 공개하지 않았다면 AI 생태계에서 덜 주목을 받거나 많은 투자금을 확보하지 못했을 수도 있다. 실제 chatGPT 출시 이후 후속 모델을 비공개로 전환하면서 엘런 머스크 같은 일부 투자자들은 더 이상 투자하지 않기로 하였다. 그리고 2025년에 라마, 딥시크, 큐웬 등의 오픈소스 모델들이 영향력을 확대하며 gpt-oss를 공개하며 기존 오픈소스 모델들을 견제하고 있다.

이들 외에도 많은 기업들이 단순히 생태계 기여라는 표면상 목적이 아닌 실질

적 공개 목적은 경쟁 기업 견제, 기술 격차 해소, 자사 기술 우수성 홍보, 생태계 영향력 확대 등의 전략적 목적을 가지고 있음을 유추할 수 있다.

2) 중국 기업의 차별화 오픈소스 전략 : 적극적 공개(혁신 기술 + 오픈소스 라이선스)

2025년 들어서 딥시크, 알리바바 등의 중국 기업들이 자체 AI 기술을 미국 기업보다 더욱 적극적으로 공개하며 큰 주목을 받고 있다. 이들의 적극적 공개 전략의 배경에는 미국 기업 중심의 AI 생태계에서 중국 기술력을 적극적으로 입증하고 영향력을 확대하기 위해서 이다.

이미 미국 기업들의 오픈소스 모델들이 AI 생태계에서 우월한 지위를 차지하고 있기 때문에 이러한 패러다임을 깨기 위해 우수한 혁신 기술(비용 절감)과 더 개방적인 오픈소스 라이선스를 채택해 공개하고 있다. 실제로 딥시크의 R1 모델은 MoE, 지식 증류 등 다양한 기술을 채택하여 학습 비용을 1/18 수준으로 낮추며 큰 주목을 받았다. 이는 AI 생태계에 큰 영향(엔디비아 주가 하락, AI 서비스 비용 하락 등)을 주며 딥시크에서 기술·산업적 영향력을 확보하게 해주었다.

알리바바도 우수한 성능의 큐웬을 허깅페이스에 공개하며 큰 주목을 받고 있다. 큐웬은 명령어 수행, 논리적 추론, 독해, 수학 등 다양한 면에서 우수한 성능을 보이며 최고 수준의 성능을 보이고 있다. 그리고 아파치 라이선스로 공개하며 모델 사용자의 저작권 우려는 해소하며 빠르게 영향력을 확대하고 있다.

이와 같이 중국 기업들은 미국 기업 보다 더 적극적인 공개(혁신 기술 + 오픈소스 라이선스)를 통해 기술 우수성과 저작권 침해 위험 해결이라는 차별화 전략을 통해 오픈소스AI 생태계에서 영향력을 확대하고 있다.

3) 오픈소스AI 영향력 확대 방안 : 빠른 기술 혁신과 잦은 반복 출시로 개발자 락인(Lock-in)

메타가 처음 라마를 공개되던 23년에 라마 1은 구글 딥마인드의 폐쇄형 모델인 Gopher 대비 14.6개월의 기술 격차가 있었지만, 24년 라마 3이 공개될 시점에는 Claude 3 Opus 대비 2.3개월의 기술 격차를 보이며 빠른 기술 발전으로 기술 격차를 해소하였다. 이와 같이 빠른 기술 혁신을 통해 폐쇄형 상용AI 모델과의 격차를 해소하면서 많은 개발자와 연구자의 지속적인 관심을 받으면서 관련 생태계를 지속적으로 확장할 수 있던 배경이 되었다.

25년에는 중국 기업들이 2-3개월 주기의 반복적인 잦은 모델 출시로 오픈소스 AI 기술 주도권을 확대해 나가고 있다. 딥시크는 2023년 11월 첫 모델 출시 이후 V2(2024.1), V3(2025.1), R1(2025.1), R1-0528(2025.5)을 연이어 공개하며 평균 3.2개월 주기로 시장에 기술 혁신을 가속화하고 있으며, 알리바바도 큐웬 2.0(2023.10), 큐웬2.5(2024.9), 큐웬 3(2025.5) 으로 이어지는 잦은 업데이트로 성능 향상과 개발자의 지속적인 관심을 받고 있다.

이러한 현상은 아직 AI 기술의 완성형 기술이 아니기 때문에 지속적인 기술 혁신이 가능하기 때문이다. 따라서, 제품/서비스 출시 후 안정화와 성능 향상을 위해 사후 관리가 중요하듯이 오픈소스AI 생태계에서도 오픈소스 모델의 지속적인 성능 향상을 위한 잦은 업데이트가 필요하다는 것이다. 중국 기업들은 기술 혁신 가속화와 연계한 잦은 반복 출시를 통해 최근 미국 기업들의 오픈소스 생태계 영향력을 추월하고 있다는 이야기가 나오고 있다.

4) 다양한 오픈소스AI 공개 방안 : 완전 공개, 제한적 공개, 폐쇄 전략

글로벌 기업들의 오픈소스 모델 공개 현황을 보면, 다양한 공개 방안들이 활용되고 있다. 이 공개 방안들은 기업의 전략적 위치에 따라 선택적으로 활용되고 있기 때문에 오픈소스AI 전략을 위한 수행 방안으로 볼 수 있다.

첫 번째 방안은 완전 공개 방안으로 딥시크, 알라바바, 미스트랄과 같이 상대적으로 생태계 추격 기업들의 전략이다. 아파치 라이선스, MIT 라이선스와 같이 오픈소스 라이선스를 채택함으로 사용자에게 저작권 침해 부담을 줄여주면서 실질적인 오픈소스AI를 공개하는 방안이다.

두 번째 방안은 제한적 공개 방안으로 메타의 라마가 이 방안을 채택하고 있다. 메타가 이 방안을 선택할 당시에는 우수한 성능을 가진 LLM이 없던 시기이었다. OpenAI의 chatGPT가 AI 서비스 시장을 장악해 나가며 chatGPT의 높은 비용이 이슈가 되었기 때문에 라마의 제한적 공개에도 많은 개발자과 연구자들은 열광하며 라마를 활용하면서 파생 모델 수 6만 5천개 이상의 강력한 생태계를 확보해 나갈 수 있었다. 하지만, 최근 중국 기업들이 오픈소스 라이선스로 모델을 공개하면서 우월적 지위에 도전을 받고 있다.

세 번째 방안은 단계적 폐쇄 전략이다. 이는 OpenAI가 취했던 방안으로 chatGPT 서비스 출시 이전에는 모델(GPT-1, GPT-2)을 공개하며 기술 영향력을

확대하였으나 경쟁 기술이 없는 상황에서 모델을 비공개하며 본격적인 수익화를 추구하였다. 이러한 방안은 엘런 머스크가 투자한 xAI도 Grok을 일시적으로 공개하였다가 비공개한 사례가 있으며, 메타도 중국 기업들의 모델이 공개되기 전에 향후 모델 비공개를 검토하고 있다는 이야기가 있었다.

5) 오픈소스AI 수익화 모델 : 간접 수익화 - 자체 제품/서비스 혁신 및 로스 리더(Loss Leader) 전략

메타, 구글, 알리바바, 딥시크 등은 오픈소스 모델 공개를 단순히 AI 생태계의 기술적 기여에 머무르지 않고, 공개된 오픈소스 모델의 개방형 기술 검증을 통한 사업화에 적극 나서고 있다. 이들 기업들은 오픈소스 모델을 자체 플랫폼(클라우드, 검색, 쇼핑 등)에 적용하여 제품/서비스 혁신을 통해 매출 확대를 추진하고 오픈소스 모델과 연계된 상용AI 서비스를 제공하며 수익을 창출하고 있다. 특히 공개된 오픈소스 모델의 우수성을 인정받게 되면 해당 상용AI 서비스의 사용자 수가 급증하며 영향력을 확대되고 있다. 또한, 우수한 오픈소스 모델을 기반으로 개발 편의성 및 안전성이 향상된 상용AI 모델 공급 및 관련 솔루션 제공 등을 추진하고 있다.

이와 같이 오픈소스 모델은 그 자체 만으로 직접적인 수익 창출이 불가능하지만, 생태계와 연계한 기술 혁신을 통한 자체 제품/서비스 혁신을 통한 매출 증대, 경쟁력 있는 오픈소스 모델을 통한 상용AI 솔루션 제공을 통해 간접적 수익 창출에 기여하고 있다. 특히 오픈소스 모델과 연계한 상용AI 솔루션 공급은 기존 오픈소스 수익화 모델인 오픈코어 방식과 같으며 이는 전통적인 상품 마케팅에서는 로스 리더(Loss-Leader) 전략 혹은 미끼상품 전략이라고 한다. 실제로 구글은 제미나이 서비스와 경량형 오픈소스 모델인 젬마와 연계해서 고객을 확대하고 있으며, 딥시크, 알리바바 등도 오픈소스 모델과 이를 활용한 AI 서비스를 제공하면서 고객 확대를 통해 수익화를 추구하고 있다.

6) 오픈소스AI의 주요 수익원 : B2B 시장(엔터프라이즈 고객)

오픈소스AI 기업의 주요 수익원으로 B2B 시장이 주목받고 있다. 구글의 경우 버라이즌, 벨 캐나다, 베스트바이 등이 제미나이 기반 고객 지원 서비스를 구축하였다. 알리바바는 큐웬을 기반으로 90,000개 기업 고객[235]을 확보했다고 밝히고

있으며, 대표적으로 Snowflake(데이터 분석), BNP Paribas(규제 문서 분석 70% 시간 단축), 시스코(네트워크 자동화) 등이 있다. 딥시크는 중국 자동차 미스트랄I AI는 오렌지(€50M 비용 절감), 르노(개발 기간 6개월 단축)와 계약하며 유럽 AI 시장의 40%를 점유하고 있다. 국내 기업인 업스테이지는 삼성전자(문서 처리 70% 단축), 신한은행(10만 달러 절감) 등 300개 고객사[236]를 확보하고 있다.

이러한 딥시크, 라마, 큐웬 등의 오픈소스 모델 기반 상용AI 서비스의 경쟁력은 비용 효율성으로 최대 chatGPT 대비 1/60이라는 낮은 가격으로 주목을 받고 있다. chatGPT의 높은 가격이 부담스러운 기업들은 오픈소스 모델로 성능이 검증된 오픈소스AI 기업의 상용 서비스를 활용하거나 자체 AI 솔루션을 구축함으로써 비용 부담을 낮추고 있다. 이러한 경향이 확산되고 있으며, 스탠포드대학의 AI 인덱스 보고서는 이러한 트렌드를 기반으로 AI 산업 구조가 자본 경쟁에서 효율성 경쟁으로 변화하고 있다고 지적하고 있다.

235) Alibaba Cloud (Alizila), 2025, "Alibaba Introduces Qwen3, Setting New Benchmark", https://www.alizila.com/alibaba-introduces-qwen3-setting-new-benchmark-in-open-source-ai-with-hybrid-reasoning/

236) 업스테이지. (2025). Solar: Apache 2.0 open-source strategy and B2B success. 업스테이지 Official Report. https://www.업스테이지.ai/ and https://huggingface.co/업스테이지

제4장 국내외 오픈소스AI 현황 분석

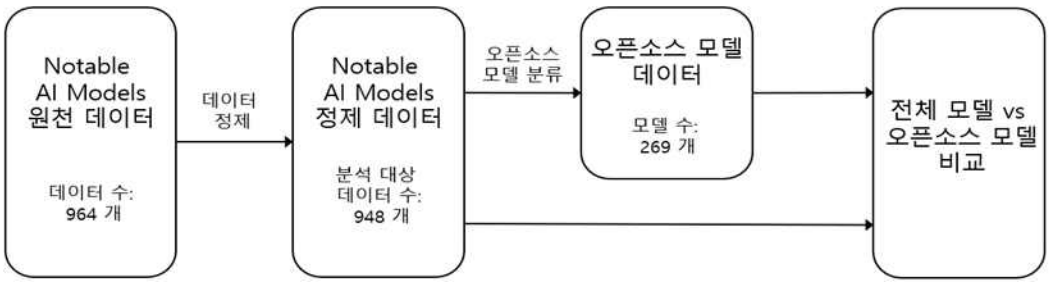
제1절 글로벌 오픈소스 모델 현황 분석

1. 분석 개요 및 방법

글로벌 오픈소스AI 현황 분석을 위해 전세계 유명 AI 모델 정보를 누적 관리하는 EpochAI의 Notable AI Models(유명 AI 모델들) 데이터를 가지고 분석하였다. EpochAI는 글로벌 AI 동향(기술, 기업 등)을 전문으로 연구하는 연구기관으로 OpenAI, 구글 딥마인드, 영국 과학혁신기술부와 협력 연구를 수행하는 미국 소재의 비영리 기관이다[237]. 유명 AI 모델들 자료는 1950년부터 2025년 현재(본 보고서 기준 ’25년 6월 초)까지 공개된 글로벌 AI 모델 중에 4가지 조건(① 공인 벤치마크의 최첨단 개선, ② 인용 수 1000개 이상, ③ 기술 발전의 중요성, ④ 중요한 활용)중 하나 이상을 충족한 965개의 AI 모델 정보를 제공하는 자료이다.

본 연구는 유명 AI 모델 자료에서 OSI의 오픈소스AI 정의(OSAID)에서 규정된 주요 공개 항목(데이터, 모델 구조, 웨이트, 학습 & 추론 코드 등)에 해당되는 필드를 중심으로 오픈소스 모델을 분류하고 AI 모델 중심의 글로벌 오픈소스AI 현황을 분석하고자 한다. 하지만, 유명 AI 모델 자료는 해당 정보를 모두 제공하지 않고 학습 데이터 정보(Training Dataset), 모델 접근성(Model Accessibliity), 학습 코드 접근성(Training Code Accessibility), 추론 코드 접근성(Inference Code Accessibility) 정보만 제공하기 때문에 이들 4개 필드를 기반으로 269개의 오픈소스 모델을 선정하였다.

[그림 24] 글로벌 오픈소스 모델 현황 분석 개요



(자체 작성)

237) EpochAI, https://epoch.ai/, 2025.12.27. 방문.

그리고, 오픈소스 모델 선정 과정에서 일부 데이터의 오류가 발견되어 사전 데이터 정제 작업을 거쳤으며 그 과정 중에 17개의 데이터를 삭제하여 분석에 활용된 AI 모델 수는 총 948개가 되었다. 그리고, 유명 AI 모델 자료에서 결측치가 많은 필드를 제외하고 데이터 분석 정보의 유용성을 감안하여 아래 표와 같은 필드들을 선정하여 분석하였다. 선정된 필드명과 그 의미, 필드별 결측치를 제외한 분석 가능 모델 수, 해당 필드를 활용한 분석 결과를 아래 표로 정리하였다. 그리고, 분석 결과는 전체 모델과 오픈소스 모델로 선정된 모델 간 비교할 수 있도록 전체 모델 현황과 오픈소스 모델 현황을 같이 제공하였다.

〈표 30〉 머신러닝 시스템 수정을 위해 선호되는 형태

필드 이름(의미)	분석 가능 모델 수	분석 결과
Model (모델 명)	948	전체 데이터(모델) 수
Orgainization (참여 기관)	931	참여 기관별 개발 참여 모델 수
Country (참여기관 국가)	924	참여기관 국가별 모델 수
Orgainization Categorization (조직 분류)	924	조직 유형별 모델 수
Publication Date (발표 일시)	946	연도별 발표된 모델 수
Domain (유형)	945	유형별 모델 수
Task (활용 분야)	943	활용 분야별 모델 수
Notability Criteria (선정 기준)	935	유명 모델 선정 기준별 모델 수
Trainining Dataset (학습 데이터 정보)	534	학습 데이터셋별 활용 모델 수
Model Accessibility (모델 접근성)	592	접근성 유형 별 모델 수
Training Code Accessibility (학습 코드 접근성)	579	접근성 유형별 모델 수
Inference Code Accessibility (추론 코드 접근성)	247	접근성 유형별 모델 수

(자체 작성)

2. 기초 통계 분석 결과

1) 참여 기관(Organization) 분석

해당 데이터는 총 948개 데이터 중 결측치 17개를 제외한 931개(98.21%)의 유효 데이터를 활용하여 전체 모델과 오픈소스 모델에 참여하는 주요 참여 기관들의 현황을 비교 분석하였다. 우선 전체 모델 현황을 기반으로 높은 빈도를 보인 상위 20개 참여 기관을 대상으로 오픈소스 모델 현황과 비교 분석하였다.

분석 결과, 구글(Google)이 전체 모델 기준 가장 많은 빈도 91개이었으며, 이 중 오픈소스 모델은 19개이었다. 그리고, 구글의 하위 조직인 DeepMind(3위), 구글 Brain(4위), 구글 Research(14위) 들도 상위 20위 안에 포함되며 AI 모델 연구 개발 및 오픈소스 모델 공개를 활발히 추진하는 것을 알 수 있다.

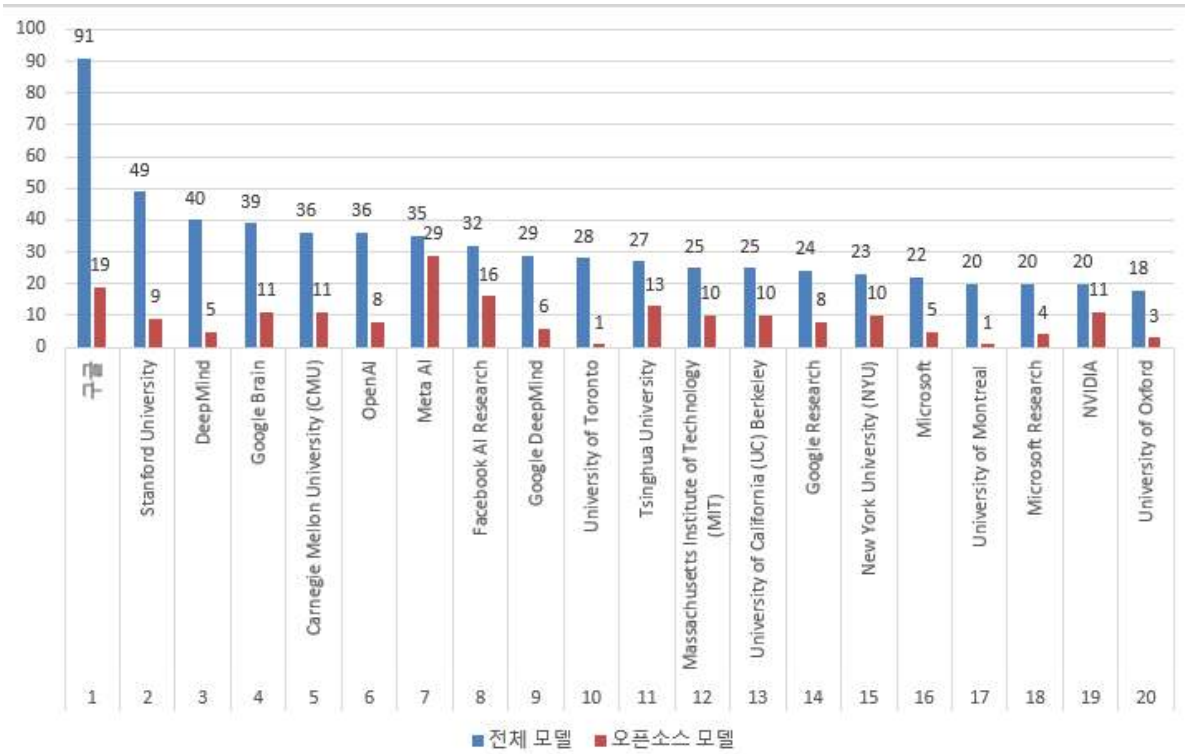
전체 모델 기준 스텐포드 대학(Stanford University, 49개), Deepmind (40개), 구글 Brain (39개), Carnegie Mellon University(36개)와 OpenAI(36개), 메타(Meta) AI(35개) 등이 상위권에 위치하였다. 전체 모델 기준 20위 권 안에는 토론토 대학(University of Toronto, 10위), 칭화 대학(Tsinghua Univsersity, 11위), 몬트리올 대학(University of Montreal, 18위), 옥스퍼드 대학(Univesity of Oxford, 20위)를 제외한 16개가 미국 기관과 대학들이었다. 이는 기존 AI 생태계를 미국이 기업과 대학들이 선도하고 있음을 보여준다.

오픈소스 모델을 기준으로 할 경우에는 메타 AI가 29개로 가장 많이 오픈소스 모델을 공개하고 있는 것으로 분석되었다. 메타 AI는 전체 모델(35개) 중 82.86%라는 매우 높은 비율로 모델 공개에 적극적이었으며 메타의 전신인 페이스북(Facebook) AI는 16개로 3위에 위치하고 있을 정도로 AI 개발에 있어 오픈소스 전략을 적극 추진하고 있음을 알 수 있다.

이어서 구글(19개), 페이스북 AI(16개), 칭화 대학(13개), 알리바바(12개), 구글 Brain(11개), CMU(11개), NVIDIA(11개) 순이었다. 또한 오픈소스 모델의 경우 상위 20위 건 안에 칭화 대학(4위), 알리바바(5위), 뮌헨 기술 대학(Technical University of Munich) 만이 비 미국 참여 기관과 대학들이었다.

하지만 칭화대와 알리바바가 전체 모델과 비교해서 높은 순위를 차지하며 중국이 오픈소스 모델 개발에 적극적으로 참여하고 있다는 근거로 볼 수 있다.

[그림 25] 참여 기관별 모델 수 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

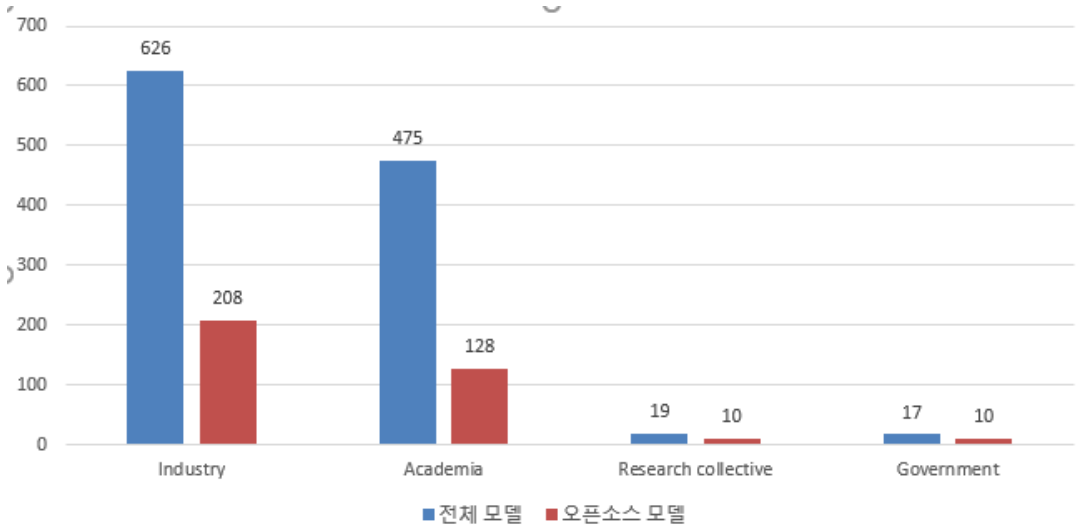
2) 기관 분류(Organization Categorization) 분석

기관 분류 데이터는 전체 데이터 948개 중 결측치가 24개로 이를 제외한 전체 데이터 수는 924개(97.5%)의 유효 데이터를 제공하고 있다. 기관 분류는 산업계(Industry), 학계(Academia), 연구계(Research Institute), 정부(Goverment)로 구분되고 있으며, 모델 개발 참여 기관별 유형 정보를 제공하고 있다.

분석 결과, 전체 모델 기준 산업계 626건(67.7%)으로 가장 많았으며 이어서 학계가 475개(51.4%), 연구기관 19개, 정부 17개 순이었다. 오픈소스 모델도 같은 순서로 산업계 208개, 학계 128개, 연구기관 10개, 정부 10개 순이었다.

이 결과에서 AI 모델 개발은 기업과 대학들이 주도하고 있음을 알 수 있으며, 그 중에서도 기업들이 미래 경쟁력 확보를 위해 적극 참여하고 있음을 알 수 있다. 그리고 연구기관과 정부는 AI 모델 개발에서 오픈소스 모델 비중이 각각 52.6%, 58.8%를 차지하며 공적 연구를 위해 적극적으로 모델을 공개하고 있다고 판단된다.

[그림 26] 기관 유형 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

3) 참여기관 국가(Country) 분석

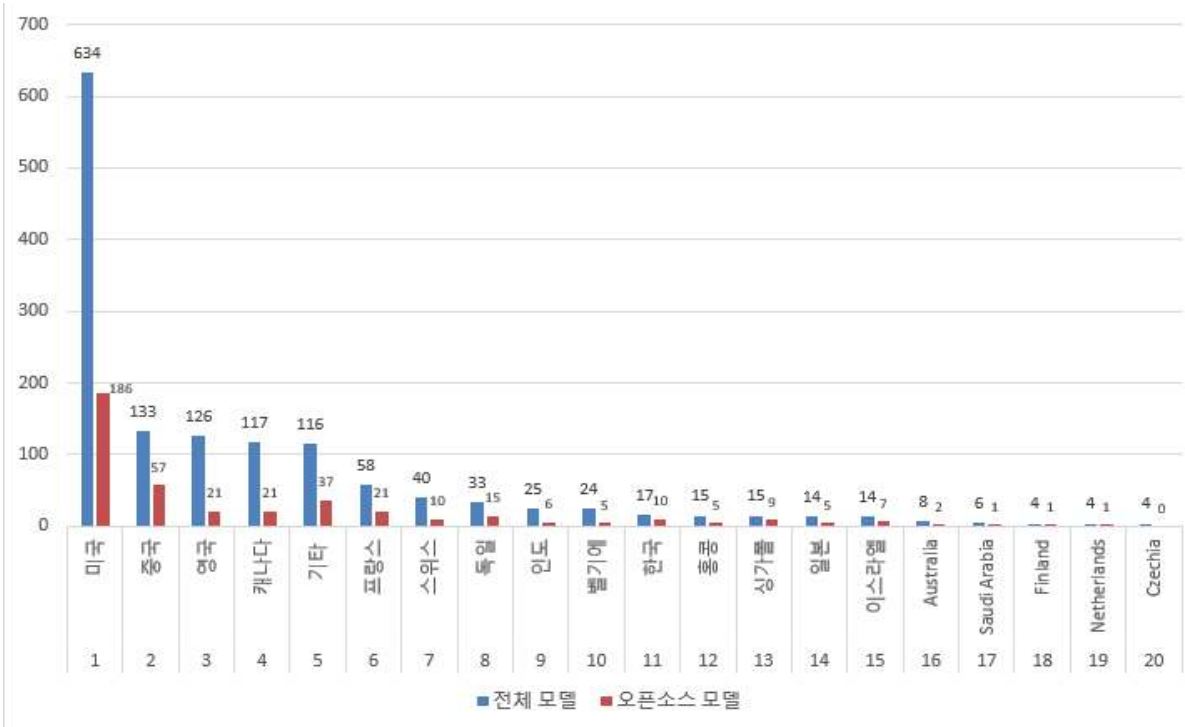
참여기관 국가 데이터는 전체 948개 데이터 중 결측치 24개를 제외한 924개(97.5%)의 유효 데이터를 제공하고 있다. 데이터는 참여 기관별 소속 국가 정보를 제공하며 분석 결과 총 32개국으로 분류되어 있고 주요국이 아닌 경우에는 Multinational로 분류하고 있다. 이 데이터는 전체 모델 기준 상위 20개국을 기준으로 전체 모델과 오픈소스 모델을 비교 분석하였다.

전체 모델 기준 미국(United States)이 가장 많은 634개로 전체 모델의 68.6%에 참여하며 가장 높은 비중을 보였으며 이어서 중국(133개), 영국(126개), 캐나다(117개), 기타(116개), 프랑스(58개), 스위스(40개) 순서 이었으며, 한국은 전체 11위로 17건의 유명 모델 개발에 참여하였다.

오픈소스 모델의 경우에도 미국이 압도적으로 많은 186개로 AI 기술 개발 선도국임을 알 수 있으며, 이어서 중국(57개), 기타(37개), 영국(21개), 캐나

다(21개), 독일(15개) 이었으며, 한국은 10개로 전체 7위를 차지하였다. 이 결과에서 중국(42.1%), 기타 국가(31.9%), 프랑스(36.2%), 독일(45.5%), 한국(58.8%), 홍콩(33.3%), 일본(35.7%)로 비영어권 주요 AI 기술 추격국들은 오픈소스 모델 개발에 상대적으로 적극 참여하고 있음을 알 수 있다.

[그림 27] 참여기관 소속 국가 현황(전체 모델 상위 20개국 vs 오픈소스 모델)



(자체 작성)

4) 발표 일시 (Publication Date) 분석

발표 일시 데이터는 전체 948개 데이터 중 결측치 2개에 불과하며 총 946개(99.79%)의 유효 데이터를 제공하며 AI 모델의 발표 시기를 의미한다. 이 데이터를 통해 유명 AI 모델은 1950년부터 발표되었음을 알 수 있으며, 분석은 모델의 발표 연도를 기준으로 전체 모델과 오픈소스 모델을 비교 분석하였다. 그래프로 시각화한 자료는 2010년 이후의 연도별 전체 모델 수와 오픈소스 모델 수를 보여준다.

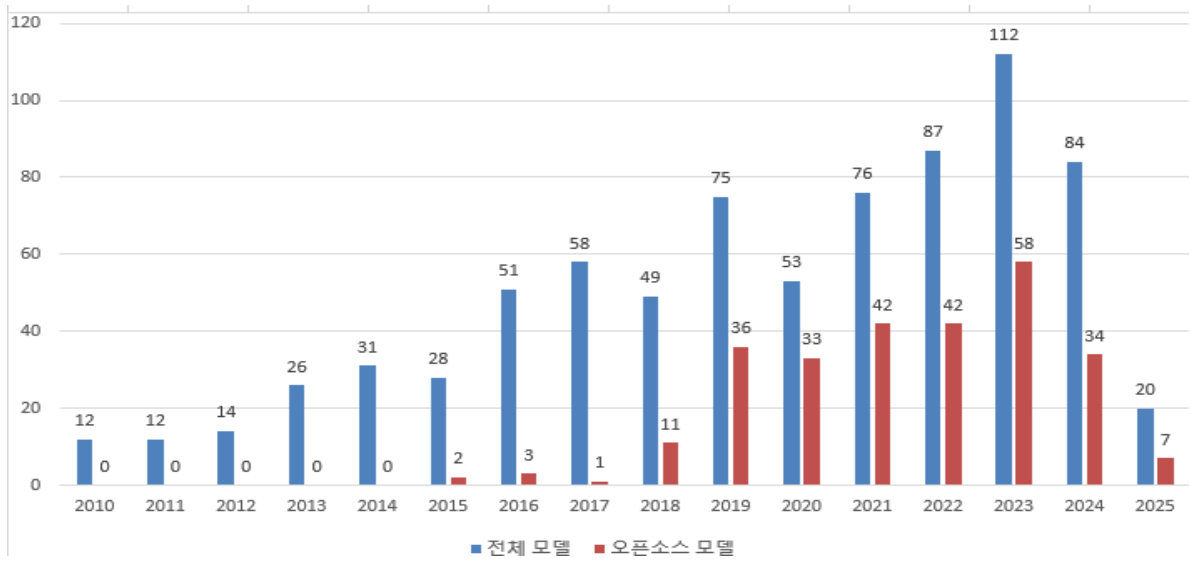
분석 결과 2004년 까지는 연도별 최대 10개에서 1개 사이로 발표되었기 때문에 2000년부터 2025년까지의 데이터를 그래프로 시각화하여 연도별 빈

도로 변화 추세를 보여주고 있다.

전체 모델의 경우, 2006년부터 10개 이상의 모델들이 안정적으로 발표되었으며, 2013년부터 발표 모델 수가 26개로 증가하며 그 이후에는 지속적으로 증가하는 추세를 보이고 있다. 가장 많은 유명 AI 모델 수는 2023년 기준 112개이었다.

오픈소스 모델의 경우 2014년까지는 유명 모델이 없었으며, 2015년부터 1개 ~ 3개로 나타나기 시작하였으며 2018년에 11개 증가하며 그 이후 지속적으로 증가하는 추세를 보이고 있다. 가장 많은 유명 오픈소스 모델 수가 많은 해는 2023년으로 58개로 전체 모델의 51.79%가 오픈소스 모델일 정도로 크게 큰 비중을 차지하였다. 오픈소스 모델의 비중이 가장 큰 해는 2020년으로 62.26%ˆ이었으며, 2019년부터 유명 모델 중 오픈소스 모델 비중이 50% 안팎을 보일 정도로 오픈소스 모델의 영향력이 확대되었음을 알 수 있다.

[그림 28] 연도별 모델 발표 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

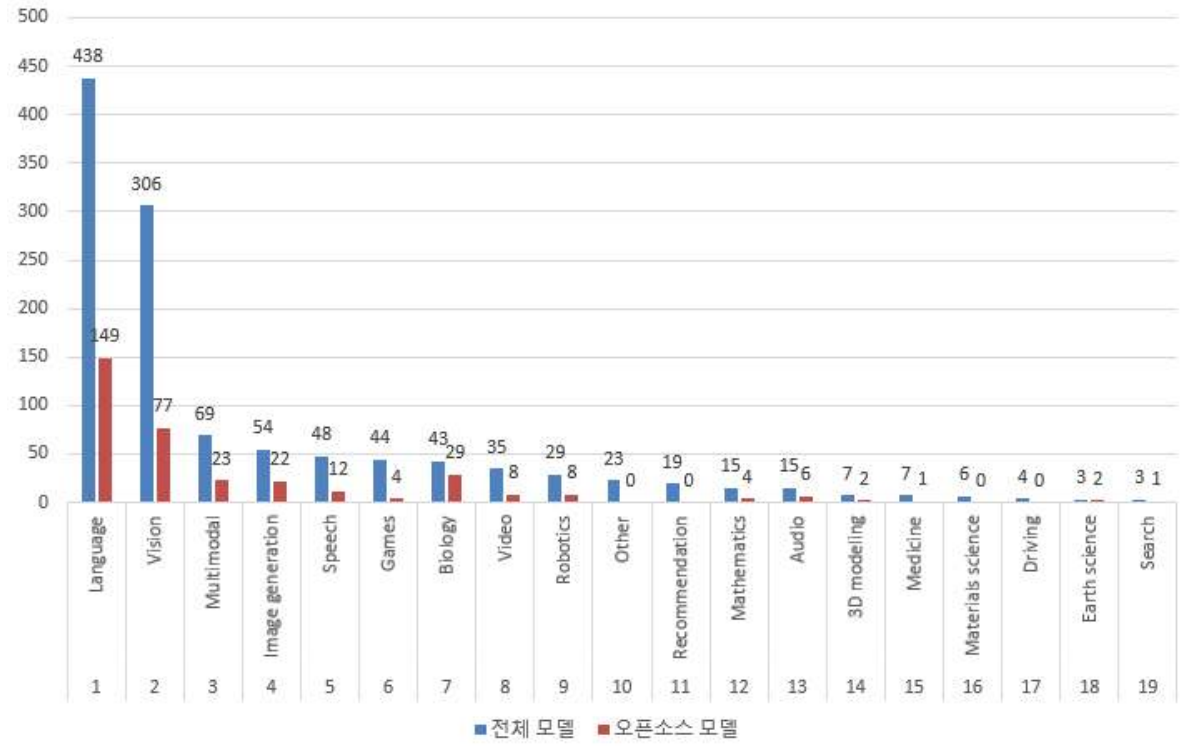
연도별 모델 발표 현황 분석 결과에서 전체 모델과 오픈소스 모델 모두 2024년과 2025년에서 감소하고 있는데, 이는 유명 모델의 선정 기준이 까다롭기 때문에 발표되었지만 유명 모델로 선정되지 않은 모델이 있을 것으로 판단된다.

5) 모델 유형(Domain) 분석

모델 유형 데이터는 전체 948개 데이터 중 결측치 3개를 제외한 945개(99.68%)의 유효 데이터를 제공하고 있다. 분석 결과 총 19개의 모델 유형이 도출되었으며 전체 모델과 오픈소스 모델을 비교하여 분석하였다.

전체 모델 기준으로 언어(Language) 모델이 438개로 가장 많이 개발되었으며, 이어서 비젼(Vision) 모델(306개), 멀티모달(Multimodal) 모델(69개), 이미지 생성(Image Generation) 모델(54개), 음성 발표(Speech) 모델(48개) 순시었다.

[그림 29] 모델 유형 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

오픈소스 모델의 경우에는 언어 모델이 148개로 가장 많았으며, 이는 최근 생성AI 모델에서 문서 요약, 생성, 번역 등에 널리 활용되는 언어 모델이 각광받고 있는 동향을 반영하고 있었다. 이어서 비젼 모델(77개), 생물학(Biology) 모델(29개), 멀티모달(23개), 이미지 생성 모델(22개) 순이었다. 오픈소스 모델에서 생물학 모델이 높은 순위를 보이는 것은 과학계의 연구

목적으로 개발된 모델이 많이 공개되었기 때문으로 추정된다. 이는 참여 기관 유형에서 학계가 2번째로 많이 차지한 결과와 관련있다.

또한 오픈소스 모델 비중이 높은 모델 유형은 생물학 모델(67.4%), 이미지 생성 모델(40.74%), 언어 모델(33.8%), 멀티모달 모델(33.3%) 순이었다. 이 결과에서 생물학 모델을 제외한 이미지 생성, 언어, 멀티모달 분야는 상용AI 서비스에서 주로 제공되는 모델들로 오픈소스 모델이 상용AI의 대안으로 전략적으로 개발되고 있는 현실을 보여준다.

6) 활용 분야 (Task) 분석

활용 분야 데이터는 전체 948개 데이터 중 결측치 5개를 제외한 943개(99.47%)의 유효 데이터를 제공하고 있다. 이 데이터는 개발된 AI 모델이 어떠한 분야에서 활용되는지에 관한 데이터로 분석 결과 총 136개의 활용 분야가 도출되었으며 전체 모델 기준 상위 21개 활용 분야(공동 19위가 3개임)에 대해 오픈소스 모델과 비교하여 분석하였다.

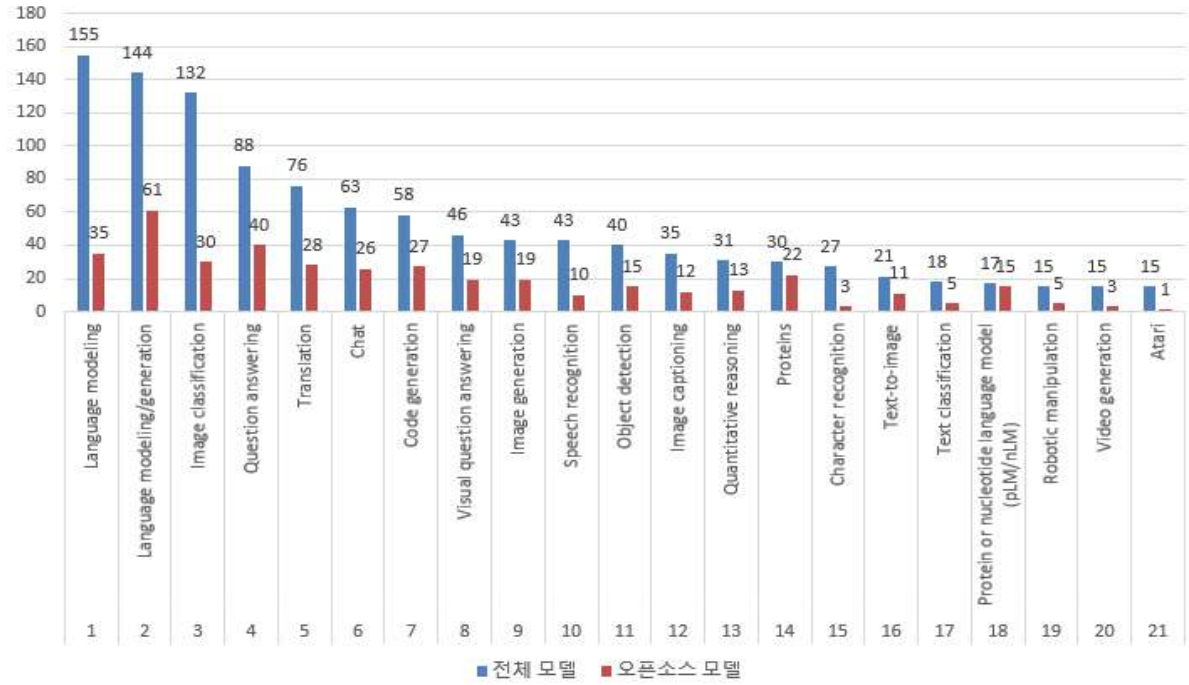
전체 모델을 기준으로 가장 많은 활용되는 분야는 언어 모델링(Language modeling) 분야로 총 155개의 모델들이 있었으며, 이어서 이미지 분류(Image classification, 132개), 질문 응답(Question Answering, 88개), 번역(Translation, 76개), 대화(Chat, 63개) 순이었다. 이들 분야는 최근 생성AI 시장에서 가장 각광받는 분야들이다. 이 결과에서 언어 모델링과 언어 모델링/생성 분야가 가장 높은 빈도를 보여주며 유명 모델의 가장 많은 활용 분야가 자연어 처리 분야임을 알 수 있다.

오픈소스 모델의 경우는 언어 모델링/생성(Language modeling/generation)이 61개로 가장 많았으며, 이어서 질의 응답(40개), 언어 모델링(35개), 이미지 분류(30개), 번역(28개), 코드 생성(27개)로 전체 모델과 유사한 순서로 개발 되고 있음을 알 수 있다. 그리고 분야별 오픈소스 모델 비중이 높은 분야는 단백질 언어 분야(88.2%), 단백질 분야(73.3%)로 생물학의 두 분야가 가장 높게 나왔으며, 이어서 문서/이미지 전환(Text-to-image) 분야(52.4%), 질의 응답(45.5%), 이미지 생성(44.2%), 언어 모델링/생성(41.7%)로 높게 나왔다. 이 결과는 생물학 분야에서 오픈소스 모델이 많이 개발되고 있음을 알 수 있으며, 생성AI 분야에 활용할 수 있는 유명 오픈소스 모델이 다수 있음

을 알 수 있다.

다만, 이 데이터 분석에 있어 유사 분야가 세분화되어 있어서 언어 모델링과 언어 모델링/생성과 같이 중복 분야를 세부적으로 구분하고 있어서 분석시 참고할 필요가 있다.

[그림 30] 모델 활용 분야 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

9) 유명 모델 선정 기준(Notability Criteria) 분석

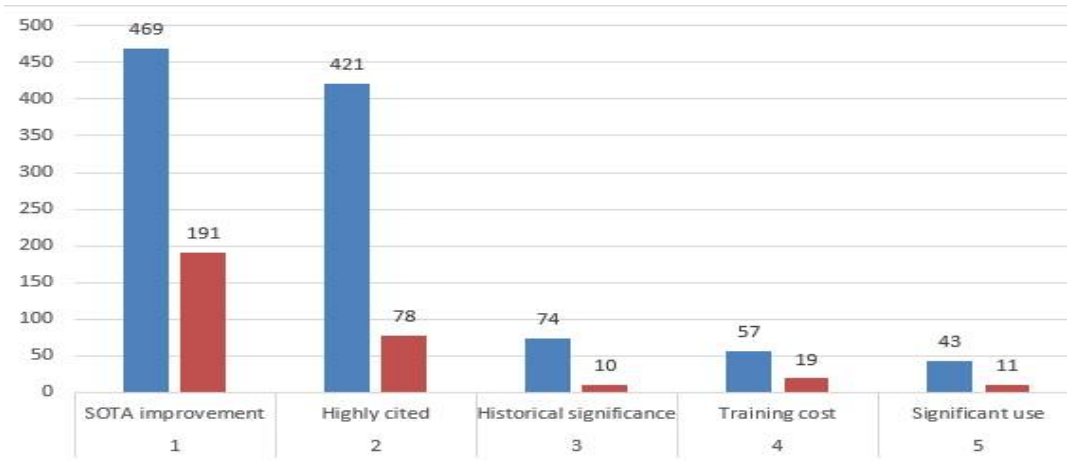
유명 모델 선정 기준 데이터는 전체 946개 데이터 중에서 13개의 결측치가 있어서 총 935개(98.63%)의 유효 데이터를 제공하고 있다. EpochAI의 유명 AI 모델 목록의 선정 기준(이유)을 제시하는 데이터로 결과 값은 총 5개이다. 유명 AI 모델은 4가지 기준 (SOTA improvement(최신 성능 향상), Highly cited(높은 인용 수), Historical siginificance(역사적 중요성), Significant use(중요한 활용 사례)으로 선정된다고 하였으나, 분석 결과에서는 Training Cost(학습 비용)이 추가로 있었다. 비록 Training cose(학습 비용)이 현재 선정 기준으로 공개되어 있지 않으나, 향후에 추가 기준이 될 가능성이 있다. 이 데이터는 각각의 기준들에 해당되는 모델 현황을 전체

모델과 오픈소스 모델을 비교 분석하였다.

전체 모델 기준 최신 성능 향상이 469개로 가장 많았으며, 유명 AI 모델의 가장 중요한 선정 기준으로 AI 모델 성능의 중요성을 보여준다. 그리고 이어서 높은 인용(421개), 역사적 중요성(74개), 학습 비용(57개), 중요한 활용 사례(43개)로 분석되었다. 높은 인용은 논문 인용 수를 의미하며 AI 기술 분야가 학계의 연구 성과를 기반으로 기술 발전한 현황을 보여준다.

오픈소스 모델의 경우도 최신 성능 향상이 가장 많은 101개로 가장 많으며 모델 성능이 가장 중요한 선정 기준으로 판단되다. 이어서 높은 인용(79개), 학습 비용(19개), 중요한 활용 사례(11개), 역사적 중요성(10개)로 분석되었다. 그리고, 전체 모델 대비 오픈소스 모델 비중을 보면 최신 성능 향상이 40.7%로 가장 높았으며, 이어서 학습 비용이 31.6%이었고, 중요한 활용 사례가 25.6%이었다. 이는 오픈소스 모델의 중요한 고려 사항이 성능과 함께 학습 비용임을 보여준다.

[그림 31] 유명 모델 선정 기준 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

10) 학습 데이터 정보(Training Dataset) 분석

학습 데이터 정보 데이터는 전체 948개 데이터 중 결측치 414개를 제외한 534개(56.33%)의 유효 데이터를 제공하고 있다. 학습 데이터는 AI 학습에 활용되는 데이터로 모델 성능과 직결되는 중요한 정보이며 OSI의 오픈소스AI

정의 및 리눅스 재단의 모델 개방성 프레임워크 모두 모델 재현성을 위해 공개 항목으로 되어 있다. 하지만, 많은 기업들은 학습 데이터 정보를 투명하게 공개하지 않으며 오픈소스 모델로 홍보하는 경우가 종종 있다.

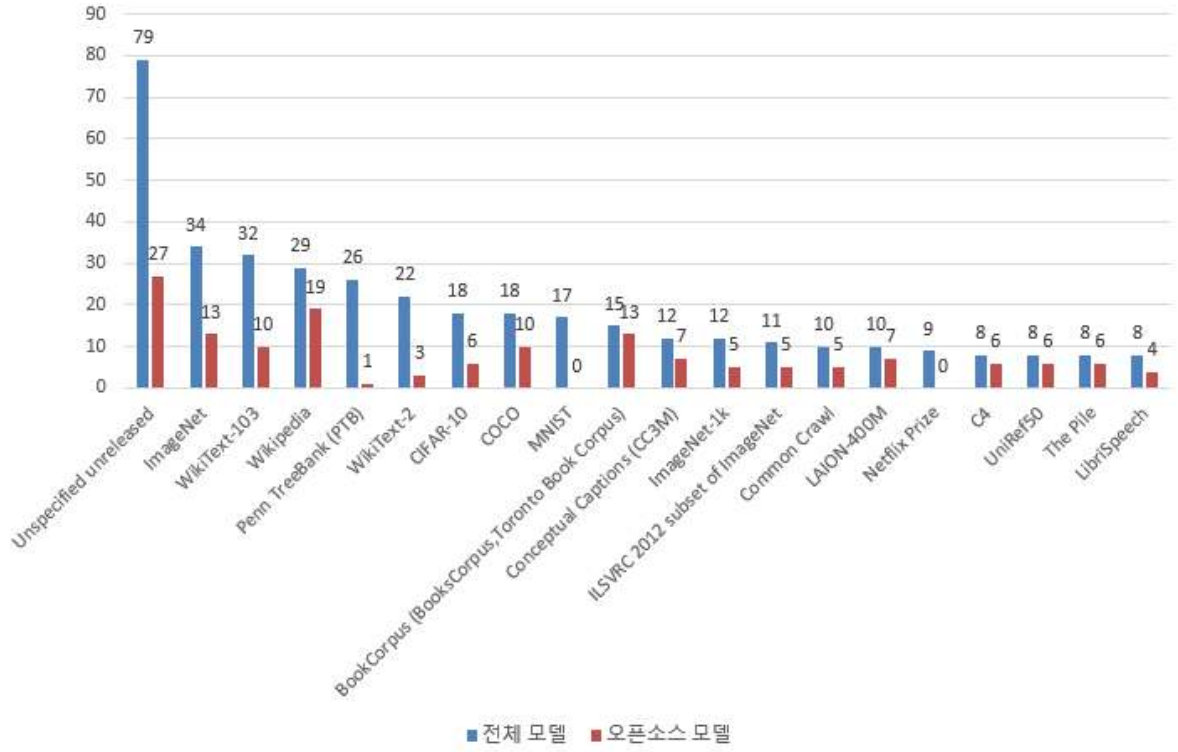
본 연구에서 오픈소스 모델 판단 기준에 학습 데이터 정보 제공 유무로만 판단한다. 그 이유는 실제 모델 학습에 활용된 모든 데이터 정보를 확인하기 어렵기 때문에 해당 필드의 데이터가 존재하는 것을 학습 데이터 정보 공개로 간주하였다. 분석 결과, 총 학습 데이터 유형은 276개 데이터들이 도출되었으며, 전체 집단을 기준으로 빈도가 높은 상위 20개 항목을 선정하여 오픈소스 모델과 비교하여 분석하였다.

전체 집단을 기준으로 불특정 비공개(Unspecified unreleased)가 79개로 가장 많은 비중을 차지하였다. 이는 학습 데이터를 공개하지 않는 것을 의미하며 이 데이터 값을 가진 모델은 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크의 조건 상 오픈소스 모델이 될 수 없다. 보편적으로 학습 데이터를 공개하지 않는 이유는 모델 재현성에 제약을 두어서 기술 확보를 어렵게 하거나 비공개 데이터일 경우에 정보를 공개하지 않는다.

불특정 비공개에 이어서 가장 많이 활용된 데이터는 ImageNet(34개), WikiText-103(32개), Wikipedia(29개), Penn TreeBank (PTB, 26개), WikiText-2(22개) 순이었으며,해당 데이터들은 공개되어 있어 AI 연구용으로 자주 활용되는 데이터들이다. 이들은 비상업용 연구 목적으로 활용할 경우에 문제가 없지만, 상업적 목적일 경우에는 저작권 침해 요소에 대한 검증이 필요하다.

오픈소스 모델의 경우, 가장 많이 언급된 데이터 유형은 불특정 비공개로 이는 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 아직 보편화되지 않은 상황을 의미한다. 이러한 분석이 나온 이유는 앞서 오픈소스 모델 분류 과정에서 현실적으로 통용되는 오픈 웨이트 모델의 기준을 적용하여 모델 웨이트(Model Accessibility)를 공개한 모델들을 포함하였기 때문이다. 불특정 비공개에 이어서 많이 활용된 데이터셋은 Wikipedia(19개), ImageNet(13개), BookCorpus(13개), WikiText-103(10개), COCO(10개) 순이었다. 전체 모델 대비 오픈소스 모델의 비중을 보면 BookCorpus가 86.7%(13/15)로 가장 높고, 이어서 Wikipedia가 65.5%(19/29), CC3M이 58.3%(7/12), COCO가 55.56%(55.6%) 순으로 전체 모델과 오픈소스 모델 모두에서 공통적으로 많이 학습되는 데이터셋이었다.

[그림 32] 학습 데이터셋 현황(전체 모델 vs 오픈소스 모델)



(자체 작성)

11) 모델 접근성(Model Accessibility) 분석

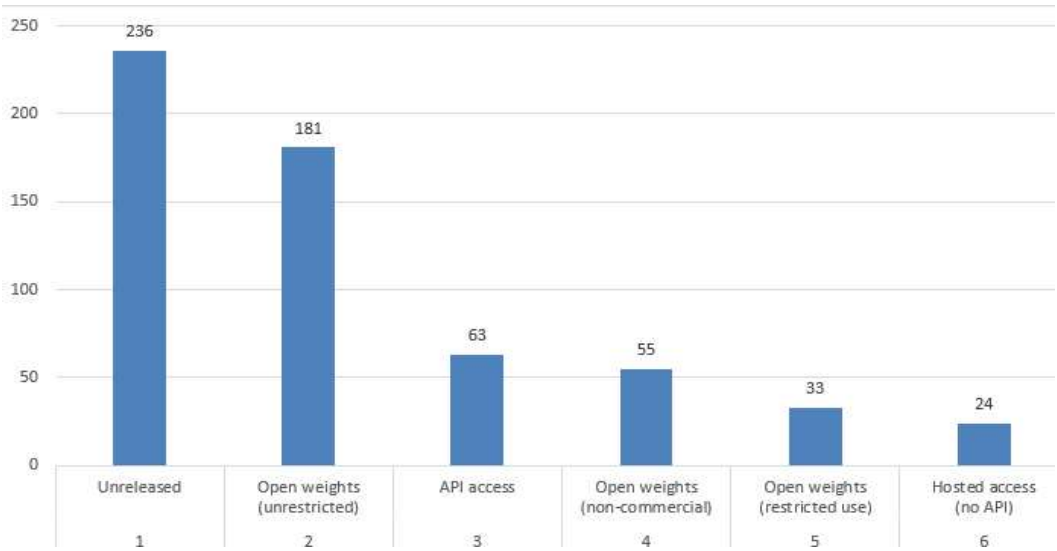
모델 접근성 데이터는 전체 948개 데이터 중 결측치 356개를 제외한 592개(62.44%)의 유효 데이터를 제공하고 있다. 이 데이터는 오픈소스 모델 여부 및 모델 개방성 평가의 핵심 요소와 관련있다. 실제로 OSI의 오픈소스AI 정의에 따른 오픈소스 모델과 오픈 웨이트의 핵심 공개 항목이며 리눅스 재단의 모델 개방성 프레임워크의 기본 공개 항목이며 원칙적으로 제한없는 사용을 허가해야 한다고 규정되어 있다.

분석 결과 이 데이터는 비공개(Unreleased), 무제한 가중치 공개(Open weights(unrestricted)), 비상업적 가중치 공개(Open weights(non-commercial)), 제한적 사용 가중치 공개(Open weights(restricted use)), API 접근(API access), 호스트 접근(Hosted access(no API))로 구분되어 있다. 그리고, 비공개가 가장 많은 236개(39.4%)로 가장 높은 비중을 차지하였다.

이어서 무제한 가중치 공개가 181개로 엄격한 오픈소스AI 정의를 충족하는 모델들이었다. 그리고, 다음은 API 접근으로 상용AI 서비스 이용 수단으로 접근성을 부여하는 모델이 63개(10.6%)이었다. 다음으로는 통상적인 오픈 웨이트 조건에 해당되는 비상업적 가중치 공개가 55개(9.3%) 이었으며, 제한적 사용 가중치 공개가 33개(5.6%)이었다.

본 연구에서 비상업적 가중치 공개 모델과 제한적 사용 가중치 공개 모델들도 오픈소스 모델로 포함시켜서 분석하고 있다. 그 이유는 OSI의 오픈소스AI 정의와 리눅스 재단의 모델 개방성 프레임워크가 광범위하게 받아들여지지 않는 현실을 감안하여 통상적인 오픈소스 모델인 오픈 웨이트 모델을 기준을 본 연구의 오픈소스 모델 선정 기준으로 하였다. 따라서 모델 접근성에 일부 제약(비상업적 활용, 제한적 활용)을 고려하지 않았다. 실제로 유명 오픈소스 모델로 알려진 메타의 라마 등은 웨이트를 공개하지만 일부 제한적 사용을 부여하고 있다.

[그림 33] 모델 접근성 현황



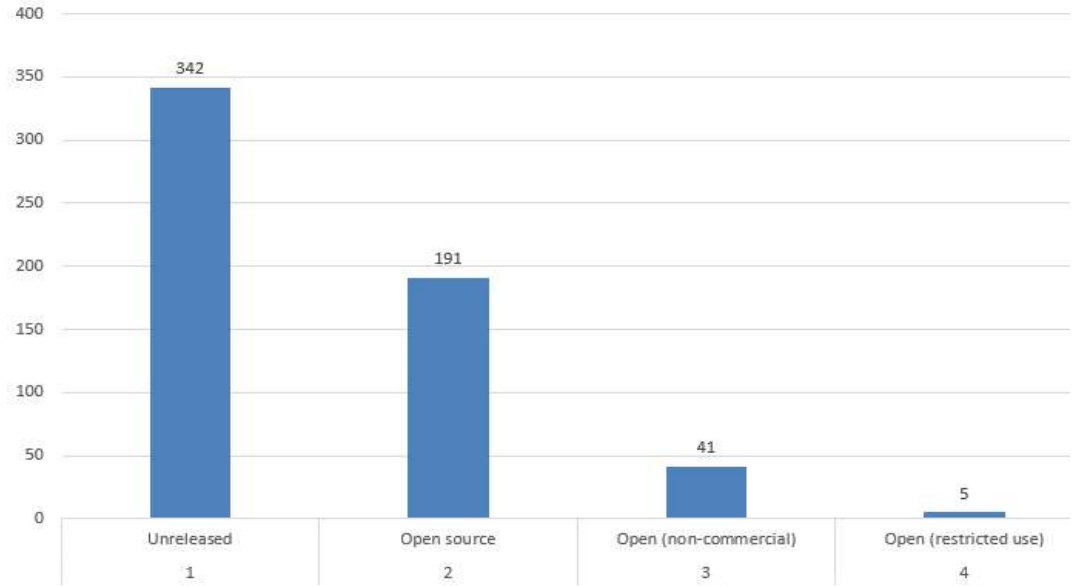
(자체 작성)

12) 학습 코드 접근성(Training Code Accessibility) 기초 분석

학습 코드 접근성 데이터는 전체 948개 데이터 중 결측치 369개를 제외한 579개(61.07%)의 유효 데이터를 제공하고 있다. 학습 코드는 모델 개발을

위해 데이터 학습을 위한 코드로 AI 시스템 재현성을 위해 필수 항목이다. OSI의 오픈소스AI 정의에서는 AI 시스템의 구성요소로 구분하고 있다. 이 데이터는 모델의 학습 코드의 접근성과 개방성을 비공개(Unreleased(비공개 모델), 오픈소스(Open source), 비상업적 공개(Open (non-commercial)), 제한적 사용 공개(Open (restricted use))로 4가지 유형으로 구분하였다.

[그림 34] 학습 코드 접근성 현황



(자체 작성)

분석 결과, 비공개가 가장 많은 342개(50.06%)이어서 많은 모델들이 학습 코드를 공개하고 있지 않음을 알 수 있다. 이는 후발 주자들을 견제하기 위한 수단으로 학습 코드를 공개하지 않고 있음을 알 수 있다. 그리고 다음으로 많은 유형은 오픈소스로 191개(32.98%)로 두 번째로 높은 비중을 차지하였다. 이어서 비상업적 공개가 41건(7.08%)이었고, 제한적 사용 공개가 5개(0.86%)로 가장 적었다. 이러한 결과는 학습 코드 공개가 모델 공개보다 제한적으로 이루어짐을 알 수 있으며, 아직까지는 오픈소스AI 생태계의 활성화 노력이 더욱 필요하다고 보여진다.

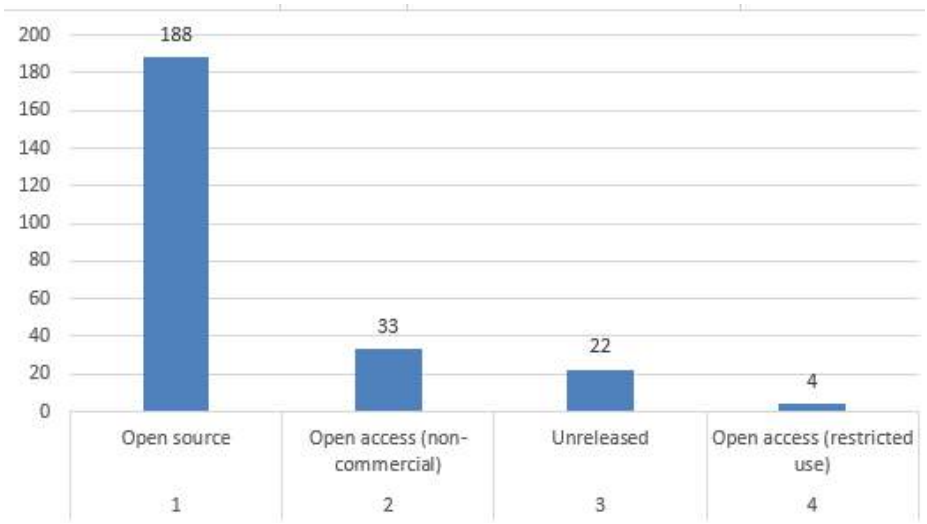
13) 추론 코드 접근성(Inference Code Accessibility) 분석

추론 코드 접근성 데이터는 전체 948개 데이터 중 결측치 701개를 제외한 247개(26.05%)의 유효 데이터를 제공하고 있다. 추론 코드는 모델을 실행한 추론 결과를 얻기 위한 SW 코드로 OSI의 오픈소스AI 정의에서는 AI 모델의 구성요소로 규정하여 오픈소스 모델이 되기 위한 필수 공개 항목이다. 이 데이터는 학습 코드와 마찬가지로 추촌 코드의 접근성과 개방성을 비공개(Unreleased(비공개 모델), 오픈소스(Open source), 비상업적 공개(Open (non-commercial)), 제한적 사용 공개(Open (restricted use))로 4가지 유형으로 구분하고 있다.

분석 결과, 오픈소스가 188개(76.11%)으로 가장 높은 비중을 보였으며, 이어서 비상업적 공개는 33개(13.36%)이었으며, 그 다음은 비공개로 22개(8.90%)이고, 마지막으로 제한적 사용 공개가 4개(1.61%)이었다.

학습 코드(369개)와 달리 결측치가 701개로 매우 많기 때문에 오픈소스 공개 비중이 다소 높아보이나 정량적 수치는 학습 코드(191개) 보다 다소 낮은 188개로 학습 코드와 추론 코드 공개가 유사하게 이루어짐을 알 수 있다.

[그림 35] 추론 코드 접근성 현황



(자체 작성)

제2절 국내 오픈소스AI 인식 및 현황 설문조사

1. 조사 개요

AI 기술 발전으로 다양한 분야에 실질적으로 적용되기 시작하면서 AI 기술 확보를 위해 오픈소스AI에 대한 관심이 크게 증가하고 있다. 따라서, 하지만, 최근 각광받고 있는 오픈소스AI에 대한 국내 인식과 현황 파악은 정책 수립을 위한 기초자료로써 중요한 의미를 가진다. 따라서, 본 연구는 오픈소스AI의 중요성과 영향력이 강해지는 시점에 새로운 오픈소스 정책 수립을 위한 국내 인식과 현황 파악을 위한 기초 자료 확보를 위한 목적으로 기획되어 추진하였다.

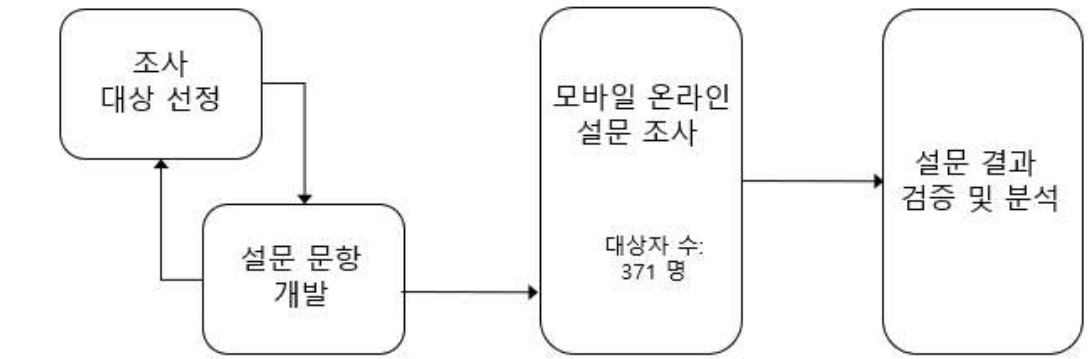
설문 조사 아래 표와 아래 그림과 같이 AI 연구개발 과정에 오픈소스AI를 활용해 본 경험이 있는 AI 개발자 371명을 대상으로 하였다. 설문 문항 개발은 설문 대상 선정과 같이 연계하여 오픈소스AI 활용 경험이 있는 AI 개발자들의 오픈소스AI 인식 및 현황에 파악하기 위한 설문 문항을 1차적으로 도출하였으며, 전문가 자문을 거쳐 최종 확정하였다. 최종 확정된 설문 문항을 기반으로 국내 최대 비즈니스 네트워크 서비스 제공사인 리멤버앤컴퍼니를 통해 모바일 온라인 조사 형태로 25년 10월 22일부터 11월 7일까지 진행하였고, 해당 설문 결과를 기반으로 기초 통계 분석을 진행하였다.

〈표 31〉 설문 조사 개요

조사 대상	AI 연구개발 과정에 오픈소스AI 활용 경험이 있는 개발자
조사 규모	371명
조사 방법	설문 문항을 활용한 모바일 온라인 조사
조사 기간	2025.10.22. ~ 11.07.
조사 기관	㈜ 리멤버앤컴퍼니
주관 기관	소프트웨어정책연구소

(자체 작성)

[그림 36] 국내 오픈소스AI 인식 및 현황 조사 방식



(자체 작성)

2. 설문 대상 선정 및 문항 개발

설문 대상 선정은 오픈소스AI에 대한 인식과 현황 조사 목적에 따라 오픈소스AI를 활용 경험이 있는 개발자로 한정하였다. 그 이유는 최근 오픈소스 AI 기술들이 AI 기술 혁신을 선도하며 AI 제품·서비스 개발에 폭넓게 활용되는 인프라적 요소가 강하기 때문이다. 실제로 리눅스 재단 보고서에 의하면 AI 개발 과정에 오픈소스 활용률이 89%에 이를 정도로 매우 높으며, 특

히 최근 오픈소스 모델들이 OpenAI의 chatGPT의 대항마로 주목받으며 빠르게 성장하고 있기 때문이다. 따라서, 실제로 오픈소스AI 기술들을 실질적으로 활용하고 있는 개발자들에 대한 인식과 현황 파악이 무엇보다 중요하다고 판단되었으며 이들이 오픈소스AI 활용 확산과 이를 통한 국가 경쟁력 확보를 위한 정책의 직·간접적인 수혜자이기 때문이다. 그러므로 설문 문항의 첫 번째 와 2번째 질문은 응답자의 적절성을 평가하기 위해 오픈소스AI 활용 경험 및 AI/SW/IT 직무를 물어보는 문항으로 구성하였다.

그리고 다음은 응답자 정보를 통한 응답자 구성의 다양화를 위해 소속 기관의 유형, 산업 분야, 소속 기관명을 문의하는 문항을 추가하였다. 이는 다양한 유형의 설문 응답을 위해 추가된 문항들로 기업 규모(대기업, 중견기업, 중소기업, 스타트업 등), 산업 분야에 따른 고른 응답자 분포와 특정 기관의 많은 응답 제한(동일 기업/기관 응답자 4명 이하)를 위해 추가하였다.

추가적인 설문 문항들은 기존 선행 문헌 조사를 통해 오픈소스AI 활용에 대한 직·간접적 주요 요인 및 저항 요인에 대한 인식과 현황에 대한 정량적 파악을 위한 40개의 설문 문항을 도출하여 Part1으로 구성하였다. Part1의 주요 내용은 오픈소스AI의 주요 활용 이유(성능 기대, 노력 기대, 사회적 영향, 촉진 조건), 저항 요인, 활용 의사, 사용 행동, 환경적 요인(기술 적합성, 환경 역동성, 제품 개발)에 대한 다양한 설문들을 수행하였다.

Part2는 오픈소스AI에 대한 개발자들의 상세 인식과 현황에 대한 구체적 설문을 통해 오픈소스AI와 관련된 정확한 현황 파악을 위해 다양한 유형들의 설문 문항들을 통해 정량적, 정성적 현상들에 대한 세부적 설문 문항들로 구성하였다. 해당 문항들을 도출함에 있어 예산적 한계로 인해 총 60개의 문항(Part 1 40개, Part 2 20개)으로 도출하였다. 도출 과정에 전문가 자문을 거쳐 설문 수정 및 추가 설문 문항들이 추가되었다.

〈표 32〉 설문 문항 개요

구성		주요 설문 내용
필터링 질문 (2개)		오픈소스AI 활용 경험, AI/SW/IT 직무 여부
기초 정보 (3개)		소속 기업/기관 유형, 산업 분야, 소속 기관명
Part1	성능 기대 (4개)	업무 속도, 업무 비용 절감, 업무 생산성 향상, 업무 성과 향상

	노력 기대 (4개)	동작 방식 이해 용이성, 능숙한 활용 용이성, 정보 취득 용이성, 유지보수 용이성
	사회적 영향 (4개)	의사결정자(상사), 동료, 고위 경영진, 사용자
	촉진 조건 (4개)	인적/재정적 지원, 필요 지식, 개발 환경/인프라, 지원 조직/전문가
	저항 요인 (5개)	저작권 침해, 개인 정보 침해, 동작 방식의 불명확성, 완전한 제어 불가, 윤리적 편향/편견
	활용 의사 (4개)	지속 사용, 추천 의향, 자주 사용, 적극 탐색
	사용 행동 (4개)	기술 문서 탐색, 새로운 정보 탐색, 학습 자료 활용, 자주 활용
	기술 적합성 (4개)	업무 성과 향상, 최선 정보 습득, 아이디어 공유, 협업 용이성
	환경 역동성 (4개)	기술 변화, 환경 변화, 취향 변화, 신제품 출시, 경쟁 변화
	제품 난이도 (2개)	제품에서 오픈소스AI 활용, 제품 개발 난이도
Part2	핵심 공개 항목 (1개)	오픈소스AI의 핵심 공개항목
	중요성 (4개)	업무 상 중요성, 국가 AI 기술력 강화, 기업 AI 경쟁력 강화, AI 기술 확산
	만족도 및 향후 예측 (2개)	오픈소스AI 활용 만족도, 활용 비중 예측
	국산 중요성 (1개)	국산 오픈소스AI 중요성
	AI 도입 유형 및 이유 (2개)	원천 AI 기술 도입 유형 및 이유
	적용 분야 및 활용 수준 (2개)	오픈소스AI 적용 분야 및 활용 수준
	모델 유형 (1개)	관심있는 오픈소스AI 모델 유형
	장점 및 단점 (2개)	오픈소스AI 활용 주요 장점 및 단점
	저해 요인 및 촉진 조건 (2개)	오픈소스AI 활용 주요 저해요인 및 촉진 조건
	플랫폼/기업 종속성 (1개)	특정 플랫폼 및 기업 종속성
	전문가 필요성 및 핵심 역량 (2개)	전문가 양성 필요성 및 전문가의 핵심 역량

(자체 작성)

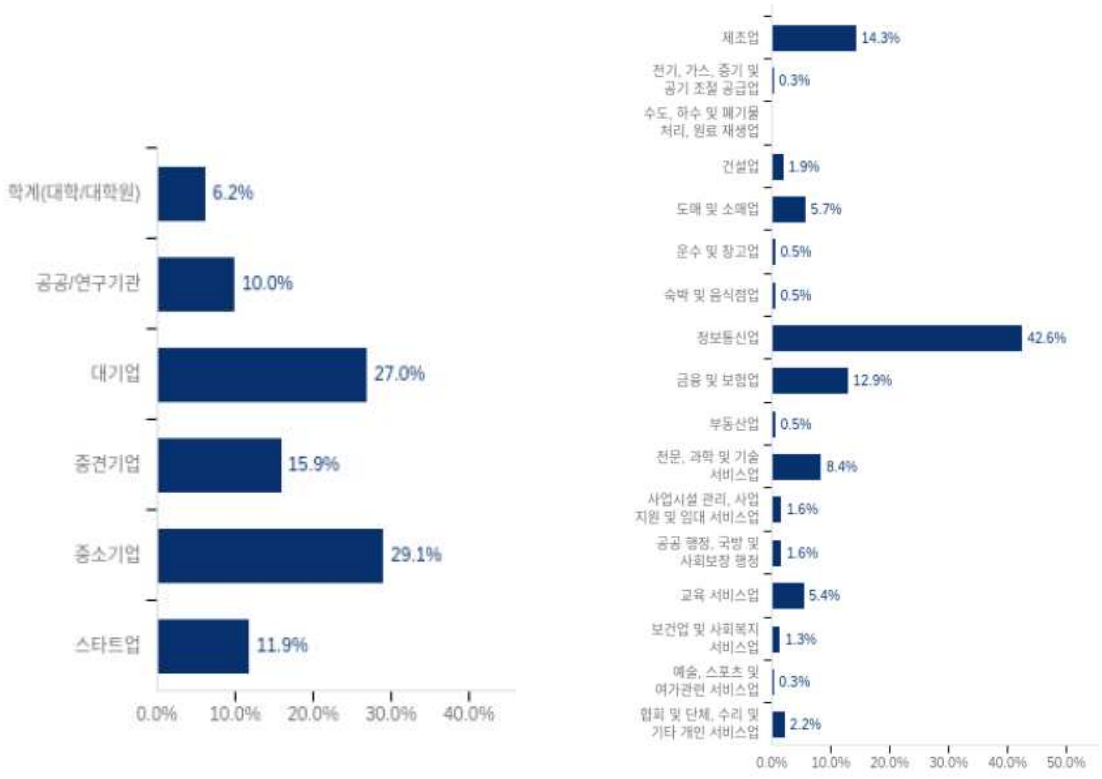
3. 설문 조사 Part 1 결과 : 오픈소스AI 활용 관련 주요 요인 분석

1) 주요 응답자 특성: 응답자 소속 기업 규모와 산업 분야

본 설문 조사 결과의 응답자 다양성을 확인하기 위해 소속 기업/기관의 유형과 기업 규모 측면를 살펴보면 중소기업 소속 응답자가 108명(29.1%)으로 가장 많았으며, 이어서 대기업이 100명(27.0%), 중견기업이 59명(15.9%), 스타트업이 44명(11.9%), 공공・연구기관이 37명(10.0%), 학계가 23명 (6.2%) 순으로 다양한 유형의 기관/기업과 다양한 규모의 기업 소속 응답자들이 설문 조사에 참여하였다.

그리고, 응답자가 속한 산업 분야들을 보면 정보통신업 소속 종사자가 158명(42.6%)로 가장 많았으며, 이어서 제조업 53명(14.3%), 금융 및 보험업 48명 (12.9%), 전문, 과학 및 기술 서비스업 31명(8.4%) 순이었으며, 이를 포함한 16개의 산업 분야의 응답자들이 설문 조사에 응답하며 AI를 활용하는 산업 분야가 매우 다양함을 알 수 있었다.

[그림 37] 주요 응답자 특성



(a) 소속 기관/기업 유형 및 규모 (b) 소속 기관/기업의 산업 분야

(자체 작성)

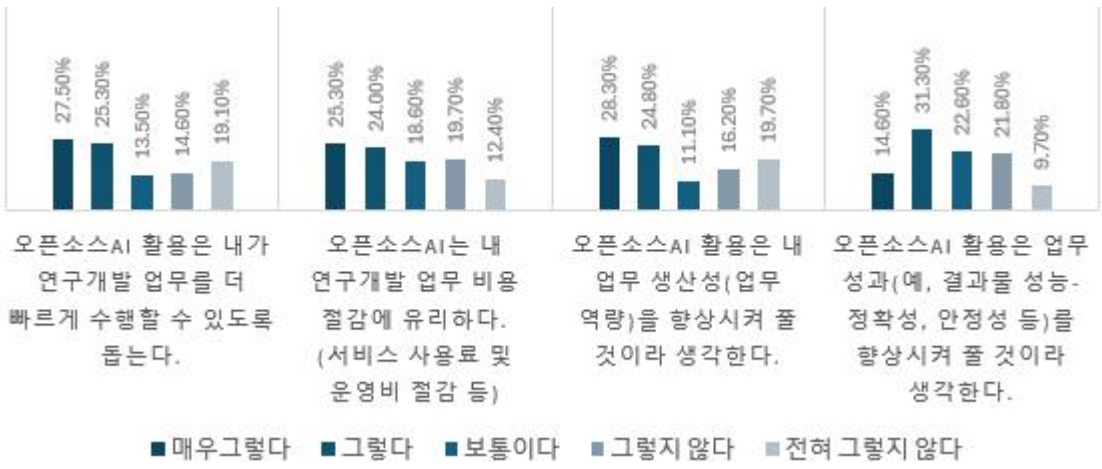
2) 오픈소스AI 성능 기대

이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI의 성능에 대한 기대를 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 개발자들이 오픈소스AI와 유사한 오픈소스에 대한 성능 기대를 기반으로 4가지 주요 특성(① 빠른 업무 수행, ② 업무 비용 절감, ③ 업무 생산성, ④ 업무 성과)에 대한 관련성을 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 업무 생산성이 가장 많은 197명(53.1%)의 응답 결과를 얻었고, 이어서 빠른 업무 수행 196명(52.8%), 업무 비용 절감 183명(49.3%), 업무 성과 170명(45.9%)로 응답하였다.

이러한 결과는 오픈소스AI 활용을 통해 업무 생산성이 가장 큰 장점으로 판단되며 업무 성과(결과물 성능-정확성, 안정성 등)은 상대적으로 기대가 낮은 걸로 판단된다. 그 이유는 오픈소스AI 기술이 빠르게 발전하고 있지만, 아직까지는 상용AI 서비스 대비 성능 다소 부족한 현실을 반영한 결과로 해석된다.

[그림 38] 오픈소스AI 성능 기대



(자체 작성)

3) 오픈소스AI 노력 기대

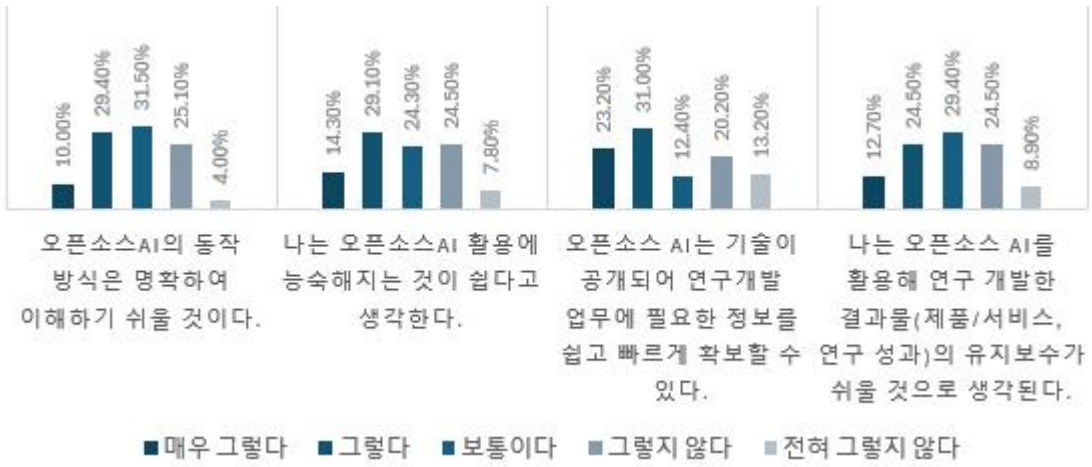
이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용에 필요한 노력에 대한 기대를 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 개발자

들이 오픈소스AI와 유사한 오픈소스 활용을 위한 노력 기대를 기반으로 4가지 주요 특성(① 동작 방식 이해, ② 능숙해지기, ③ 필요 정보 획득, ④ 유지 보수)들의 관련성을 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 필요 정보 획득이 가장 많은 201명(54.2%)의 응답 결과를 얻었고, 이어서 능숙해지기 161명(43.4%), 동작 방식 이해 146명(39.4%), 유지 보수 138명(37.2%)로 응답하였다.

이러한 결과는 오픈소스AI가 폐쇄형 상용AI 서비스보다 많은 정보가 공개되어 있어서 개발자가 정보 획득에 유리하다고 인식하고 있음을 알 수 있다. 하지만, 능숙해지기, 동작 방식 이해, 유지 보수는 상대적으로 낮은 비율의 응답자들이 동의함으로써 AI 기술적 난이도로 인해 기술 습득이 상대적으로 쉽지 않게 인식하기 있기 때문으로 해석된다.

[그림 39] 오픈소스AI 노력 기대



(자체 작성)

4) 오픈소스AI 사회적 영향

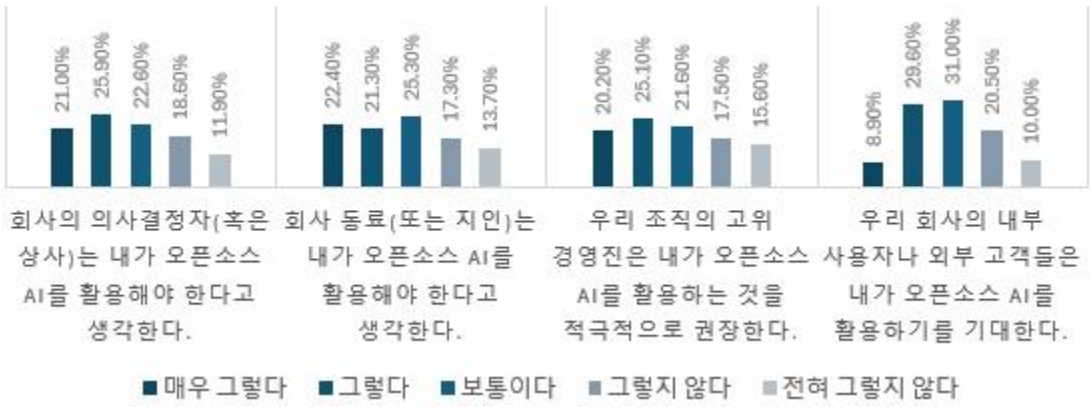
이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용시 영향을 주는 사회적 영향력을 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 개발자들이 오픈소스AI와 유사한 오픈소스 활용에 미치는 주변 인물들에 대한 4가지 주요 특성(① 의사결정자(상사), ② 동료(지인), ③ 고위 경영진, ④ 제

품/서비스 사용자)에 대한 영향력을 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 의사결정자(상사) 174명(46.9%)로부터 가장 큰 영향을 받는 응답 결과를 얻었고, 이어서 고위 경영진 168명(45.3%), 동료 162명(43.7%), 제품/서비스 사용자 143명(38.5%)로 응답하였다.

이 결과는 직장 구조 상 오픈소스AI 활용에 큰 영향을 미치는 인물은 직장 상사 혹은 고위 경영진일 수 밖에 없는 현실적 요인들이 고려된 것으로 판단된다. 그리고, 가장 영향을 미치는 제품/서비스 사용자로 나타난 이유는 사용자들은 사용된 원천 기술이 오픈소스AI인지 상용AI 서비스인지 인지하지 못하거나 관심이 없을 수 있기 때문인 것으로 해석된다.

[그림 40] 오픈소스AI 사회적 영향



(자체 작성)

5) 오픈소스AI 촉진 조건

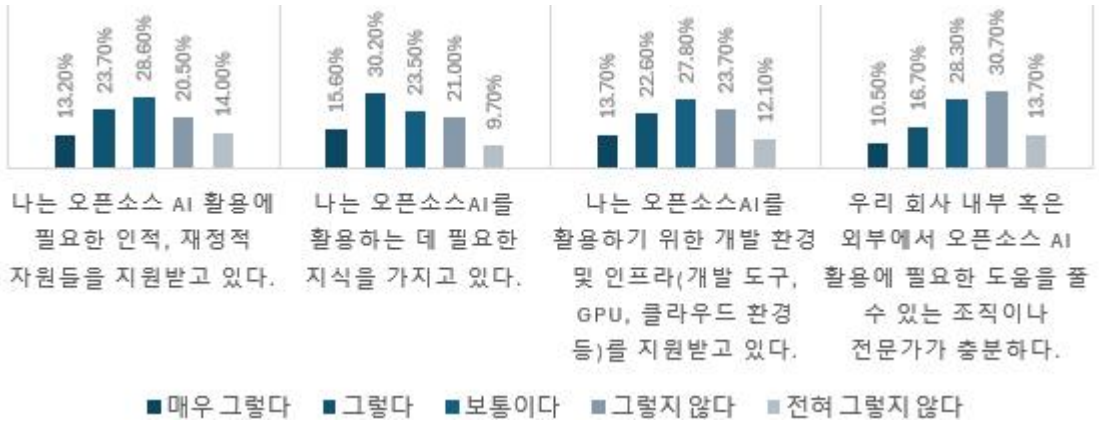
이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용을 촉진하는 다양한 조건들에 대해 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 개발자들이 오픈소스AI와 유사한 오픈소스 활용을 촉진하기 위한 4가지 주요 특성(① 인적/재정적 지원, ② 필요 지식, ③ 개발환경/인프라 지원, ④ 지원 조직/전문가)과의 관계를 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를

수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 필요 지식이 가장 많은 170명(45.8%)의 응답 결과를 얻었고, 이어서 인적/재정적 지원 137명(36.9%), 개발환경/인프라 지원 135명(36.3%), 지원 조직/전문가 101명(27.2%)로 응답하였다.

이러한 결과는 오픈소스AI 활용을 위해서는 개발자의 지식이 가장 중요한 촉진 조건으로 판단되며, 이는 후속 설문 문항에서 오픈소스AI 전문가 양성 필요성이 높게 제기되는 것과 일맥상통한다. 그리고, 많은 오픈소스AI 개발자들이 충분한 인적/재정적, 개발환경/인프라, 지원 조직/전문가가 부족하다고 인식되고 있는 것으로 판단된다. 따라서, 오픈소스AI 활용 확산을 위해서는 충분한 전문가 양성, 개발환경/인프라 지원, 전문가 조직 지원이 필요해 보인다.

[그림 41] 오픈소스AI 촉진 조건



(자체 작성)

6) 오픈소스AI 저항 요인

이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용에 저해되는 저항 요인을 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 AI 사용자들이 AI 서비스에 대한 우려 요인들에 대한 중복성을 제거하여 5가지 주요 특성(① 저작권 침해, ② 개인 정보 침해, ③ 불투명한 동작 방식, ④ 불완전한 제어, ⑤ 윤리적 편향/편견)들을 도출하여 오픈소스AI 활용에 반하는 요인들의 관련성을 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇

다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 부정적 응답(매우 그렇다와 그렇다)의 경우 저작권 침해가 가장 많은 172명(46.3%)의 응답 결과를 얻었고, 이어서 개인정보 침해 165명(44.5%), 불완전 제어 164명(44.2%), 불명확한 동작 방식 159명(42.8%), 윤리적 편향 144명(38.8%)으로 응답하였다.

이 결과는 많은 개발자들이 오픈소스AI 활용시 향후에 발생할 수 있는 저작권 및 개인정보 침해 우려를 가지고 있음을 알 수 있다. 이는 최근 오픈소스 모델로 공개되는 일부 모델들이 학습 데이터를 공개하지 않거나 일부만 공개하는 오픈 웨이트 모델 형태로 공개하기 때문으로 해석된다. 그리고, AI 신뢰성, 특히 편향에 대한 우려는 상대적으로 낮게 나왔는데, 이는 AI 결과물이 미치는 심각성이 낮기 때문으로 해석된다.

[그림 42] 오픈소스AI 저항 요인

(자체 작성)

7) 오픈소스AI 활용 의도

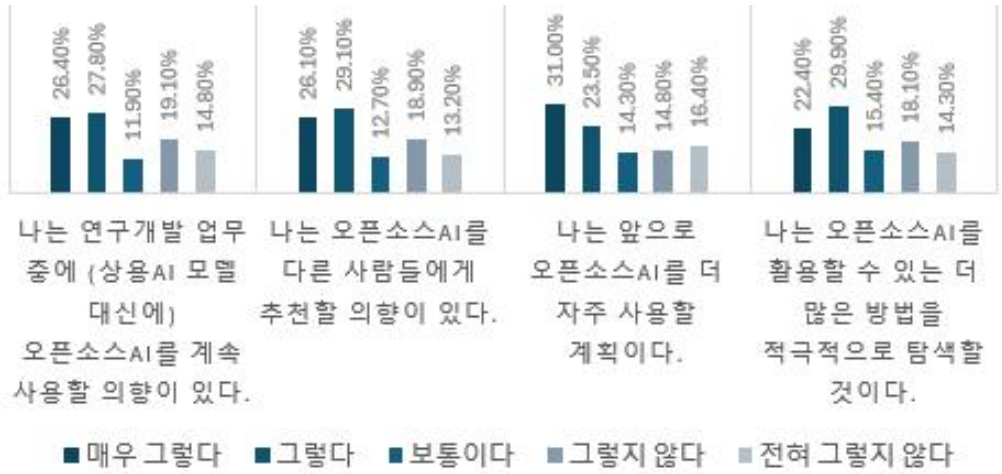
이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용에 대한 다양한 의도를 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 오픈소스AI와 유사한 오픈소스 활용에 대한 의도를 기반으로 4가지 주요 특성(① 지속 사용, ② 주변 추천 , ③ 자주 사용(빈도), ④ 다양한 방법 탐색)들에 대한 관련성을 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇

다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 주변 추천이 가장 많은 205명(55.2%)의 응답 결과를 얻었고, 이어서 자주 사용(빈도) 202명(54.5%), 지속 사용 201명(54.2%), 다양한 방법 탐색 194명(52.3%)으로 응답하였다.

이 결과는 오픈소스AI 개발자들이 오픈소스AI를 주변에 추천하고 있으며, 빈번하게 자주 사용하고, 지속적으로 계속 사용하며, 다양한 활용 방법을 탐색하고자 하는 의사가 있음을 보여준다. 그리고, 지금까지의 설문 응답 중에서 가장 긍정적 응답 결과를 보이면서 개발자들이 오픈소스AI에 대한 높은 관심을 확인할 수 있다.

[그림 43] 오픈소스AI 활용 의도



(자체 작성)

8) 오픈소스AI 실제 행동

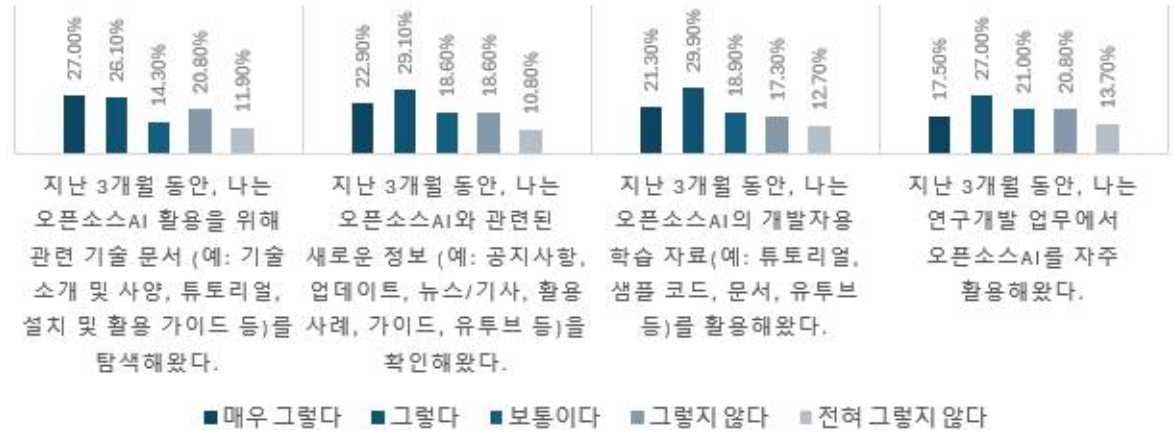
이 설문 문항들은 오픈소스AI 개발자들이 오픈소스AI 활용에 관한 실제 행동을 확인하고자 하는 문항이다. 선행 문헌 연구를 통해 오픈소스AI와 유사한 오픈소스 활용과 관련된 실제 행동을 기반으로 4가지 주요 특성(① 기술 문서 탐색, ② 새로운 정보 획득, ③ 학습 자료 활용, ④ 자주 활용)에 e 대해 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 기술

문서 탐색이 가장 많은 197명(53.1%)의 응답 결과를 얻었고, 이어서 새로운 정보 획득 193명(52.0%), 학습 자료 활용 190명(51.2%), 자주 활용 165명(44.5%)으로 응답하였다.

이 결과는 개발자들이 오픈소스AI 활용을 위해 지난 3개월간 기술 문서를 탐색하고 새로운 정보 획득과 학습 자료를 활용하여 기술 역량(지식) 획득에 많은 관심을 가지고 있음을 보여준다. 그리고, 습득한 역량을 기반으로 오픈소스AI를 자주 활용한다는 긍정적 응답이 부정적 응답보다 10%가 높았다.

[그림 44] 오픈소스AI 실제 행동



(자체 작성)

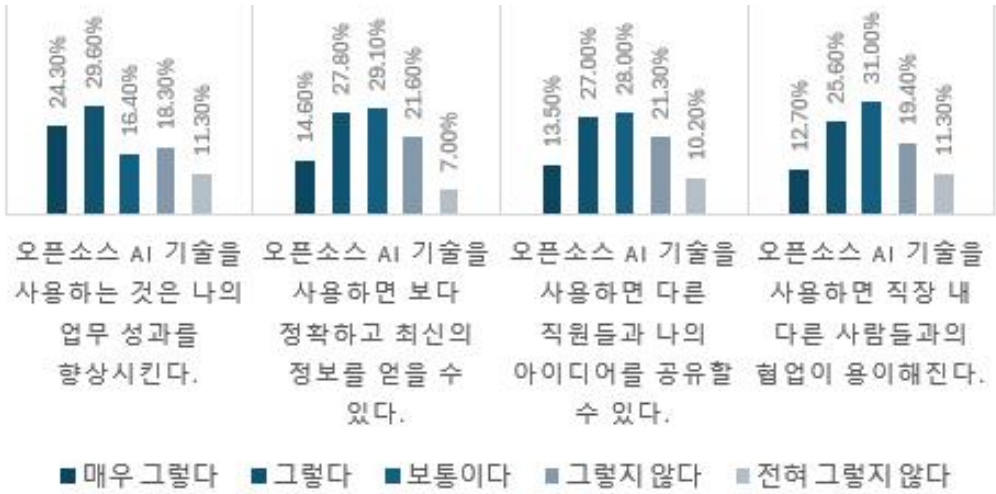
9) 오픈소스AI의 역량 적합도

이 설문 문항들은 오픈소스AI 활용 개발자들과 오픈소스AI와의 역량 적합도를 파악하고자 하는 문항이다. 선행 문헌 연구를 통해 개인과 기술과의 역량 적합도를 기반으로 4가지 주요 특성(① 업무 성과 향상, ② 최신 정보 획득, ③ 아이디어 공유, ④ 협업 용이)를 도출하여 관련성에 대해 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 업무 성과 향상이 가장 많은 200명(53.9%)의 응답 결과를 얻었고, 이어서 최신 정보 획득 157명(42.4%), 아이디어 공유 150명(40.5%), 협업 용이 142명(38.8%)으로 응답하였다.

이 결과는 개발자들이 오픈소스AI 활용의 가장 중요한 요인이 업무 성과 향상에 가장 많은 관심이 있음을 알 수 있으며, 상대적으로 최신 정보 획득과 아이디어 공유, 협업 용이는 상대적으로 낮게 응답을 하였다. 이는 개발자들이 오픈소스의 장점인 아이디어 공유와 협업 용이에 대해 상대적으로 관심이 적음을 알 수 있다.

[그림 45] 오픈소스AI의 역량 적합도



(자체 작성)

10) 오픈소스AI의 환경 변화

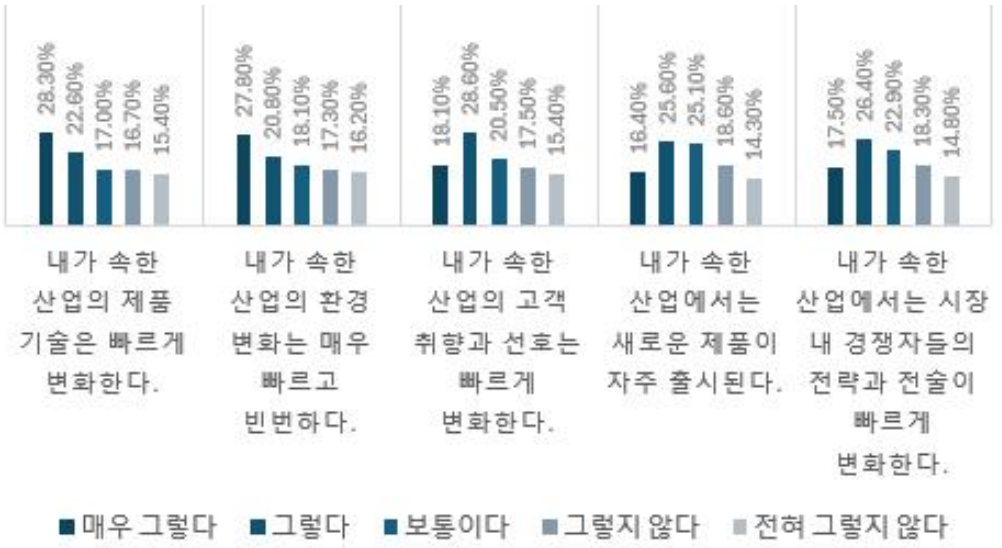
이 설문 문항들은 오픈소스AI 활용과 관련된 환경 변화를 파악하고자 하는 문항이다. 선행 문헌 연구를 통해 오픈소스AI 관련된 기술, 산업, 고객 취향 등과 관련된 주변 환경적 요소에 대한 5가지 주요 특성(① 기술 변화, ② 산업 변화, ③ 고객 취향/선호, ④ 신제품 출시, ⑤ 경쟁자 변화)과의 관계를 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 빠른 기술 변화가 가장 많은 189명(50.9%)의 응답 결과를 얻었고, 이어서 빠른 산업 변화 180명(48.6%), 빠른 고객 취향/선호 변화 173명(46.7%), 빠른 경쟁자 변화 163명(43.9%), 신제품 출시 156명(42.0%)으로 응답하였다.

이 결과는 개발자들이 오픈소스AI 활용과 관련하여 가장 큰 변화를 기술

발전 속도로 인식하고 있음을 알 수 있으며, 그리고 기술 변화에 따라 산업 변화와 고객 취향이 변화하고 있다고 인식하는 것으로 판단된다. 하지만 신제품 출시는 상대적으로 긍정적 응답이 적어 개발자들이 인식하는 신제품은 아직 상대적으로 많지 않다고 볼 수 있다. 이는 AI 산업이 최근 빠르게 성장하고 있지만 AI 거품론이 제기될 정도로 수익화에는 어려움을 겪고 있는 현실을 반영한 응답 결과로 해석된다.

[그림 46] 오픈소스AI의 환경 변화



(자체 작성)

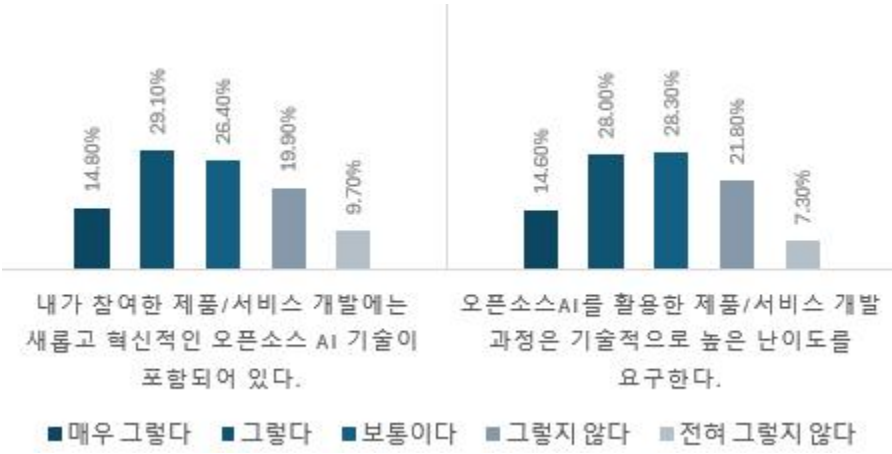
11) 오픈소스AI의 기술 복잡성

이 설문 문항들은 오픈소스AI 활용 개발자들이 오픈소스AI 기술이 실제 제품과 서비스에 반영되어 있는지와 기술적 난이도에 관련된 기술 복잡성을 파악하고자 하는 문항이다. 선행 문헌 연구를 통해 오픈소스AI 기술 복잡성과 관련된 2가지 주요 특성(① 제품/서비스 반영, ② 기술 난이도)과의 관계를 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 그렇다, 2=그렇다, 3=보통이다, 4=그렇지 않다, 5=전혀 그렇지 않다)를 사용해 응답 결과를 수집하였다. 설문 결과에서 긍정적 응답(매우 그렇다와 그렇다)의 경우 제품/서비스에 오픈소스AI가 포함되어 있다는 응답이 163명(43.9%)로 기술 난이도에 동의한 응답자 158명(42.6%)보다 다소 높다고 응답하였다.

이 결과는 개발자들이 오픈소스AI가 기술적 난이도가 높음에도 제품/서비스에 반영되는 비율이 높다고 응답한 것이다. 실제로 제품/서비스 반영에 긍정적 응답 비율이 부정적 응답 비율보다 24% 높았으며, 기술 난이도에 동의하는 응답자가 미동의하는 응답자 보다 13.5% 높았다.

[그림 47] 오픈소스AI의 기술 복잡성



(자체 작성)

4. 설문 조사 결과 Part 2 : 주요 오픈소스AI 인식 및 현황

1) 오픈소스AI 핵심 공개 항목

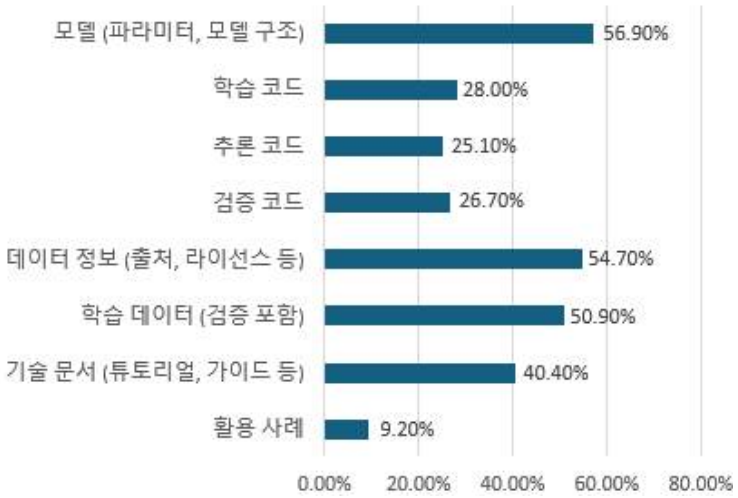
이 설문 문항은 오픈소스AI의 핵심 공개 항목에 대한 개발자들의 인식을 파악하고자 하는 문항이다. 선행 문헌(OSI 오픈소스AI 정의, 리눅스재단 모델 개방성 프레임워크)를 기반으로 모델, 학습 코드, 추론 코드, 검증 코드, 데이터 정보, 학습 데이터, 기술 문서, 활용 사례 들을 도출하여 응답자들이 최대 3개 항목을 선택할 수 있도록 하였다.

설문 결과에서 가장 많은 선택을 받은 구성 요소는 모델(파라미터, 모델 구조)로 211명(56.9%)이 선택하였으며, 이어서 데이터 정보(출처, 라이선스 등)가 203명(54.7%), 학습 데이터(검증 포함)가 189명(50.9%), 기술 문서(튜토리얼, 가이드 등)이 150명(40.4%)이 응답하였다.

이 결과는 개발자들이 오픈소스AI 활용을 위한 공개 항목의 중요도를 의미하며, 모델, 데이터 정보, 학습 데이터가 오픈소스AI 활용을 위한 가장 핵심 요소로 인식되고 있었다. 이는 단순 모델 활용 뿐만 아니라 모델 파인

튜닝 같은 최적화를 위해 데이터에 대한 수요가 높음을 알 수 있다. 그리고, 예상 외로 학습(104명), 추론(93명), 검증(99명) 코드 등은 낮은 응답률을 보였는데, 이는 기존에 공개된 학습/추론/검증 코드를 활용하여 다른 모델에 적용할 수 있다는 전문가 의견이 있었다.

[그림 48] 오픈소스AI의 핵심 공개 항목



(자체 작성)

2) 오픈소스AI의 중요성

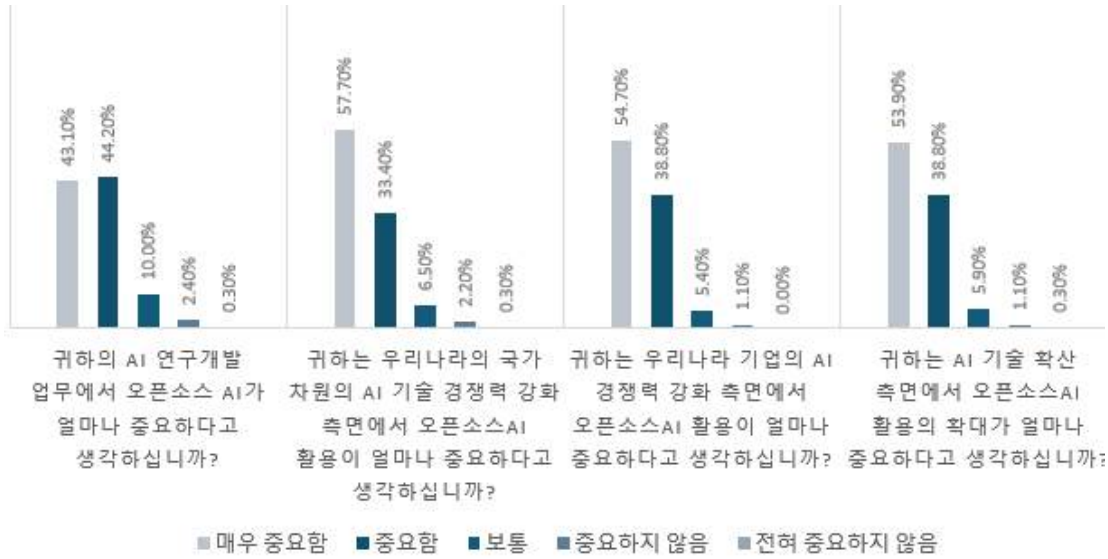
이 설문 문항들은 다양한 측면의 오픈소스AI의 중요성에 대한 개발자들의 인식을 파악하고자 하는 문항들이다. 선행 문헌과 국가 AI 정책 관점의 4가지 요소(업무 상 중요성, 국가 AI 기술력 강화, AI 기술 확산)를 도출하여 중요성를 파악하기 위한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 5점 리커트 척도(1=매우 중요함, 2=중요함, 3=보통이다, 4=중요하지 않음, 5=전혀 중요하지 않음)를 사용해 응답 결과를 수집하였다. 설문 결과에서 모든 응답 결과에서 긍정적 응답(매우 중요함과 중요함)의 비중이 최소 87%를 넘었으며, 기업 AI 경쟁력 강화 측면이 가장 많은 347명(93.5%)이 응답하였다. 이어서 AI 기술 확산 측면의 중요성이 344명(92.7%), 국가 AI 기술 경쟁력 강화가 338명(91.1%), AI 연구 개발 업무 상 중요성이 가장 낮은 324명(87.3%)로 응답하였다.

이 결과는 많은 개발자들이 다양한 측면의 오픈소스AI 중요성에 공감하고

있음을 보여주며, 특히 기업과 국가 AI 경쟁력 강화에 큰 도움이 될 것으로 인식하고 있음을 보여준다. 그리고 4개 측면의 중요도에서 부정적 인식(중요하지 않음과 전혀 중요하지 않음)은 모두 3% 이하로 응답되었다.

[그림 49] 다양한 측면의 오픈소스AI 중요성



(자체 작성)

3) 오픈소스AI 성능 만족도 및 향후 활용 예측

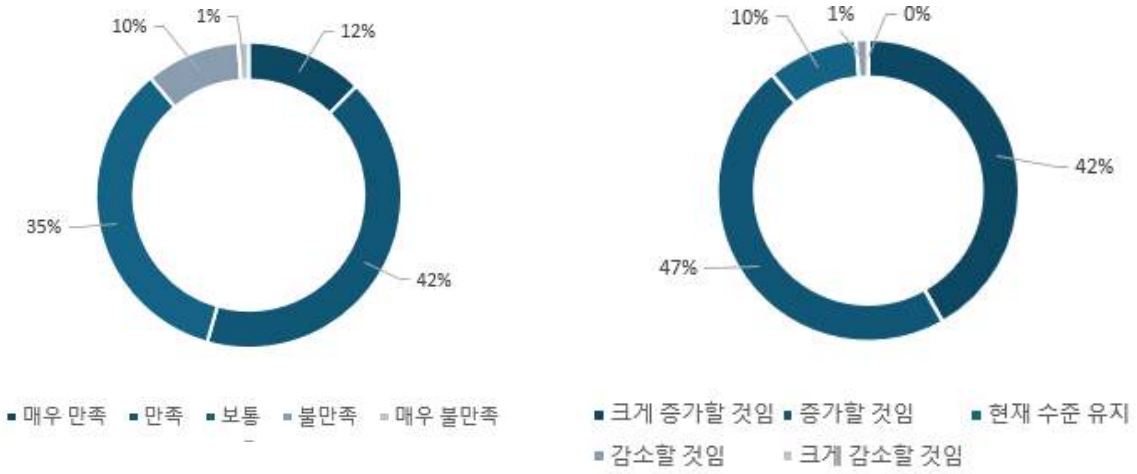
이 설문 문항들은 오픈소스AI의 성능 만족도와 향후 활용 예측에 대한 개발자들의 인식을 파악하고자 하는 문항들이다. 개발자들이 활용하고 있는 오픈소스AI의 기술적 성능에 대한 만족도와 함께 향후에 얼마나 활용이 증가할지에 대한 설문 문항을 개발하였다.

그리고 각각의 설문 문항에 대해 각각의 5점 리커트 척도(1=매우 만족/크게 증가할 것임, 2=만족/증가할 것임, 3=보통/현재 수준 유지, 4=감소할 것임, 5=크게 감소할 것임)를 사용해 응답 결과를 수집하였다. 만족도 설문 결과에서 오픈소스AI 성능 만족도에서 긍정적 응답(매우 만족과 만족)의 비중이 202명(54.4%)이었으며, 향후 활용 예측에서 긍정적 응답(크게 증가와 증가)의 비중이 330명(89.0%)이었다.

이는 개발자의 과반 이상이 오픈소스AI 기술의 성능에 만족하고 있음을 의미하며, 불만족은 41명(11.1%)에 불과할 정도로 최고 수준의 기술은 아니지만 적정 기술로써의 가치를 인정받고 있다고 판단된다. 그 결과가 개발자

의 절대 다수인 330명(89%)가 오픈소스AI 활용이 증가할 것으로 예측하였으며, 단지 5명(1.3%)의 응답자 만이 향후 활용이 감소할 것으로 예측하고 있다.

[그림 50] 오픈소스AI 만족도 및 향후 활용 예측



(a) 오픈소스AI 만족도 (b) 오픈소스AI 향후 활용 예측

(자체 작성)

4) AI 원천 기술 도입 형태와 도입 이유

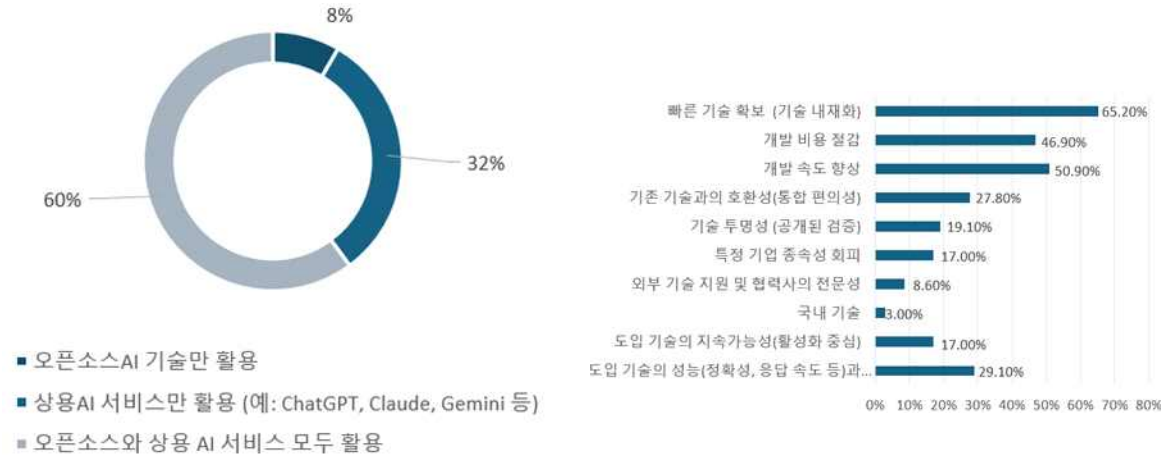
이 설문 문항들은 오픈소스AI를 포함한 AI 기술 도입 형태와 도입 이유에 대한 현황을 파악하고자 하는 문항들이다. 따라서 선행 연구 문헌(리눅스 재단 보고서)를 기반으로 개발자들이 AI 연구개발 업무에서 활용하는 AI 원천 기술(오픈소스AI, 상용AI 서비스)의 도입 형태와 그 이유에 대한 설문 문항을 개발하였다.

그리고 AI 원천 기술 도입 형태는 오픈소스AI와 상용AI로 구분하였으며, 중복 도입을 감안한 3개의 응답을 설계하였으며, 도입 이유는 10개의 응답(빠른 기술 확보, 개발 비용 절감, 개발 속도 향상, 기술 호환성, 투명성, 종속성 회피 등)으로 설계하였으며 3개의 복수 응답이 가능하도록 하였다. 도입 형태 설문 결과에서 오픈소스AI만 도입한 응답은 31명(8.4%)에 불과했으나, 오픈소스AI와 상용AI 서비스를 동시에 도입하는 비중이 가장 많은 223명(60.1%)이었다. 이는 오픈소스AI를 도입한 비중이 전체 68.5%로 약 2/3이라는 것을 의미한다.

그리고, 이러한 도입 형태를 선택한 이유로는 가장 많은 응답은 빠른 기

술 확보(기술 내재화)로 242명(65.2%)가 응답하였으며, 이어서 개발 속도 향상이 189명(50.9%)이었으며, 개발 비용 절감이 174명(46.9%), 도입 기술의 성능이 108명(29.1%), 기술 호환성이 103명(27.8%)이었다. 이러한 결과는 AI 도입시 빠른 기술 확보가 가장 중요한 요소로 판단되며 이는 최근 기술 변화가 매우 빠른 AI 생태계 특성을 반영하고 있다. 그리고 다음으로 중요한 이슈는 개발 속도 향상 및 비용 절감 같은 개발 효율성 향상이 중요한 영향을 미치고 있다고 판단된다.

[그림 51] AI 원천 기술 도입 형태와 그 이유



(a) AI 원천기술 도입 형태 (b) AI 원천 기술 도입 이유

(자체 작성)

5) 오픈소스AI 적용 분야와 활용 수준

이 설문 문항들은 오픈소스AI를 실질적으로 적용하는 분야와 활용 수준에 대한 현황을 파악하고자 하는 문항들이다. 따라서 선행 연구 문헌(리눅스 재단 보고서 등)를 기반으로 개발자들이 오픈소스AI를 활용한 AI 연구개발 결과물의 적용 분야와 활용 수준(제품화 단계)에 대한 설문 문항을 개발하였다.

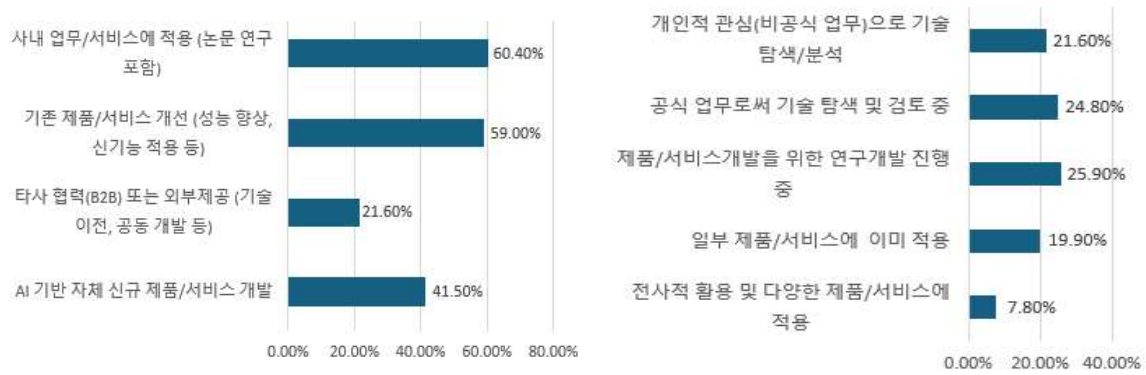
오픈소스AI 적용 분야는 기관/기업의 업무 영역을 구분하여 사내 업무/서비스 적용, 기존 제품/서비스 개선, 타사 협력 또는 외부 제공, AI 신 제품/서비스 개발 형태로 구분하며 최대 2개의 응답 결과를 선택할 수 있도록 하였다. 가장 많이 선택받은 적용 분야는 사내 업무/서비스 적용으로 224명(60.4%)가 응답하였으며, 이어서 기존 제품/서비스 개선이 219명(59.0), AI 기

반 신제품/서비스 개발이 154명(41.5%)이었다. 이러한 결과는 AI 결과물이 사내 활용 및 기존 제품 개선에 주로 활용되고 신제품/서비스 개발 혹은 AI 기술 공급은 상대적으로 적음을 알 수 있다.

이와 유사하게 오픈소스AI 기술의 활용 수준(단계) 문항은 제품화 단계(수준)을 기반으로 개인적 관심(비공식 업무), 공식적 기술 탐색 및 검토, 제품/서비스 개발을 위한 연구개발, 일부 제품/서비스 적용, 전사적 활용으로 기업 혹은 기관의 연구개발 업무의 수준을 측정하는 문항으로 설계하였다. 응답 결과는 최고 수준의 1개 문항만 선택하게 하였다. 설문 결과 응답이 고르게 분포되어 있으며 가장 많은 선택을 받은 응답은 제품/서비스 개발을 위한 연구개발 단계로 96명(25.9%)이 응답하였다. 이어서 공식 업무로써 기술 탐색/검토가 92명(24.6%)로 응답되었고, 개인적 관심이 80명(21.6%), 일부 제품/서비스 적용이 74명(19.9%), 전사적 활용은 29명(7.8%)이었다.

이 결과는 대부분의 기관과 기업에서 오픈소스AI 활용 수준이 기술 탐색 혹은 제품/서비스 개발 단계가 72.3%로 다수의 응답자들이 응답하였으며, 실제 제품/서비스를 제공하는 경우는 27.7%에 불과하다는 것을 알 수 있다. 그리고, 21.6%는 회사 업무가 아닌 개인적 관심에 의해 오픈소스AI를 활용하고 있음을 알 수 있다.

[그림 52] 오픈소스AI 적용 분야와 활용 수준



(a) 오픈소스AI 적용 분야 (b) 오픈소스AI 활용 수준

(자체 작성)

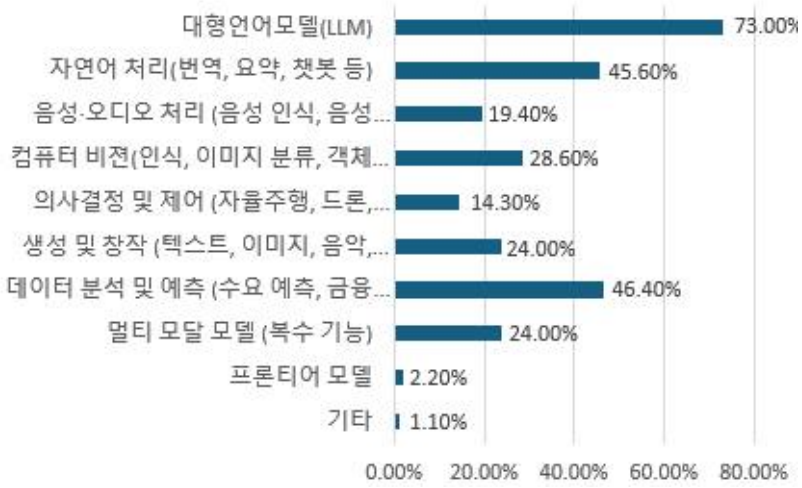
6) 관심있는 오픈소스AI 모델 유형

AI 기술이 발전해 감에 따라 활용 범위가 확대되면서 모델 유형도 다양해지고 있다. 이 설문 문항은 개발자들이 관심있어하는 오픈소스AI 모델의 유형을 파악하기 위한 문항이다. 선행 문헌(리눅스재단 보고서, EpochAI 도메

인, 전문가 자문 등)을 기반으로 주요 모델 유형과 활용 사례를 도출하여 도출하여 응답자들이 최대 3개 항목을 선택할 수 있도록 하였다.

설문 결과에서 가장 많은 선택을 받은 모델 유형은 최근 많이 주목받고 있는 대형 언어 모델(LLM)로 271명(73.0%)의 응답자가 선택하였으며, 이어서 데이터 분석 및 예측 모델이 172명(46.4%), 자연어 처리가 169명(45.6%), 컴퓨터 비전(이미지 인식 및 분류)이 106명(28.6%)로 응답하였다. 이 외에도 다양한 모델 유형들이 선택되었으며 이는 AI 활용 분야가 다양해지고 있음을 의미한다.

[그림 53] 관심있는 오픈소스AI 모델 유형



(자체 작성)

7) 오픈소스AI 장점 및 단점

이 설문 문항들은 오픈소스AI의 구체적인 장점과 단점을 파악하기 위한 문항들이다. 이미 Part1에서 성능기대, 노력기대, 저항 요인 등에 대한 설문 문항들이 있었지만 이들은 단편적 질문들이기 때문에 다양한 요인간 비교를 위해 해당 문항들을 선행 연구 문헌(리눅스재단 보고서 등)을 검토하여 설계하였으며, 최대 3개의 응답을 선택할 수 있도록 하였다.

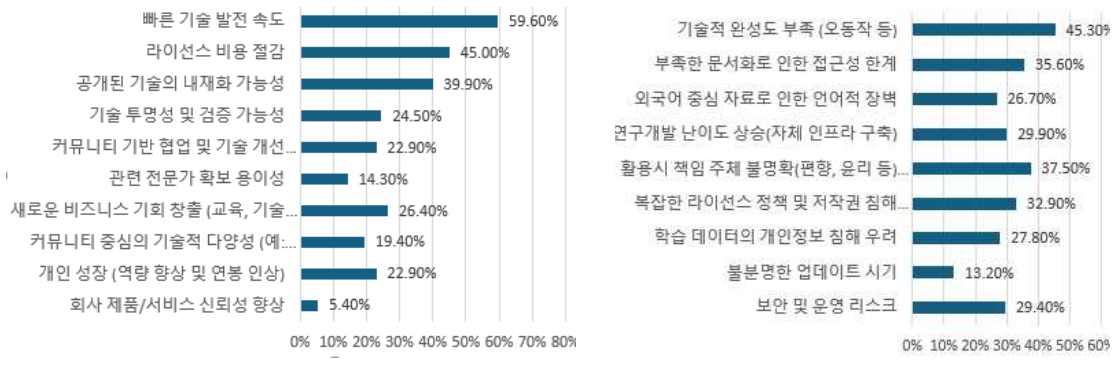
오픈소스AI 장점에 대한 설문 결과에서 가장 많은 선택을 받은 응답은 빠른 기술 발전 속도로 221명(59.6%)가 응답하였으며, 이어서 라이선스 비용 절감이 167명(45.0%), 공개 기술 내재화 가능이 148명(39.9%), 비즈니스 기회 창출이 98명(26.4%), 기술 투명성 및 검증 가능성이 91명(24.5%), 커뮤니티

기반 협업과 개인 성장이 각각 85명(22.9%)으로 응답되었다. 이 결과는 앞선 Part 1의 설문과 연계하면 오픈소스AI 선택의 핵심 요인은 기술 발전 속도와 비용 절감으로 판단되며, 더욱이 기술 내재화를 통한 비즈니스 기회 창출이 가능하다는 현실적 이유 때문이다.

오픈소스AI 단점에 대한 설문 결과는 장점과 다르게 고르게 분포되어 되어 있었다. 가장 많은 선택을 받은 응답은 기술적 완성도 부족이 168명(45.3%)이었고, 이어서 책임주체 불명확이 139명(37.5%), 부족한 문서화가 132명(35.6%), 연구개발 난이도(부담 증가)가 111명(29.9%), 보안 및 운영 리스크가 109명(29.4%), 개인정보 침해 우려가 103명(27.8%), 외국어 중심 자료(문서 등)이 99명(26.7%)로 응답되었다.

이와 같이 다양한 측면의 단점들이 제기되었으며 이는 오픈소스AI의 자발적 공개에 따른 기술적 완성도 부족 및 책임 소재 부재가 오픈소스AI 확산의 가장 큰 장애물로 판단된다. 반대로 이러한 오픈소스AI의 단점들은 오픈소스AI 활용의 애로사항으로 볼 수 있기 때문에 이러한 문제들을 해결할 수 있는 기업은 새로운 비즈니스 기회 창출이 가능하다고 볼 수 있다.

[그림 54] 오픈소스AI의 주요 장점과 단점



(a) 오픈소스AI 장점 (b) 오픈소스AI 단점

(자체 작성)

8) 오픈소스AI 활용 저해 요인 및 촉진 조건

이 설문 문항들은 오픈소스AI 활용 확산을 저해하는 실질적 요인 및 활용 촉진을 위한 실질적 요인들을 파악하기 위한 문항들이다. 이미 Part1에서 저항 요인 및 촉진 조건에 관한 설문 문항이 있었지만, 설문 문항 수 한계로 인해 다양한 측면에 대한 문항 제시를 하지 못해서 선행 연구 문헌(리눅스재단 보고서 등)을 검토하여 설계하였다. 그리고 최대 3개의 응답을 선택할 수 있도록 하였다.

오픈소스AI 활용을 저해하는 주요 요인에 대한 설문 결과에서 가장 많은 선택을 받은 응답은 조직의 오픈소스AI 전략(비전) 부재로 157명(42.3%)가 응답하였다. 이어서 인력(개발자, 기여자) 부족이 155명(41.8%), 자체 인프라 부족이 144명(38.8%), 조직의 정책/거버넌스 부재가 127명(34.2%), 개인 기술 역량 한계가 99명(26.7%), 불명확한 규제 및 위반 가능성은 94명(25.3%)로 응답되었다. 이 결과는 앞선 Part 1의 설문과 연계하면 저작권 침해, 개인정보 침해, 불완전한 동작 보다는 조직(기관/기업)의 전략/비전/정책/거버넌스 부재와 인력 부족 및 기술적 역량 부족 더 큰 저해 요인으로 판단된다.

오픈소스AI 활용을 촉진하기 위한 주요 조건에 대한 설문 결과에서 가장 많은 선택을 받은 응답은 연구개발 인프라로 203명(54.7%)가 응답하였다. 전문가 양성(기술 교육) 및 발굴이 182명(49.3%)이었으며 이어서 기업/기관 거버넌스 확립이 144명(38.8%), 오픈소스AI 활용 기업 발굴 및 지원 138명(37.2%), 커뮤니티 활성화 124명(33.4%)로 응답되었다. 이 결과는 저해 요인을 해결하기 위한 응답이 높게 나오며 일관성을 유지하는 결과로 연구개발 인프라 지원, 전문가 양성, 기업/기관 AI 거버넌스 확립, 모범 사례 발굴을 위한 기업 지원 필요성을 제기하고 있다.

[그림 55] 오픈소스AI 활용 저해 요인 및 촉진 조건

(a) 오픈소스AI 활용 저해 요인 (b) 오픈소스AI 활용 촉진 조건

(자체 작성)

9) 국산 오픈소스AI 중요성 및 특정 플랫폼/기업 종속성 우려

글로벌 오픈소스AI 생태계는 미국과 중국 기업 중심으로 빠르게 확산되고 있다보니 국내에서 많이 활용되는 오픈소스AI 기술들도 해외에서 개발되어 공개된 기술이 많다. 또한 해외 플랫폼(허깅페이스 등)을 활용하여 오픈소스AI 개발이 확산되다 보니 해외 플랫폼을 활용하는 경우가 많이 있다. 따라서, 비록 공개된 기술임에도 불구하고 해외 의존성에 대한 관심이 많이 때문에 이에 대한 AI 개발자들의 인식을 파악하고자 국산 오픈소스AI 중요성과 특정 플랫폼/기업 종속성 우려에 대한 문항들을 설계하였다.

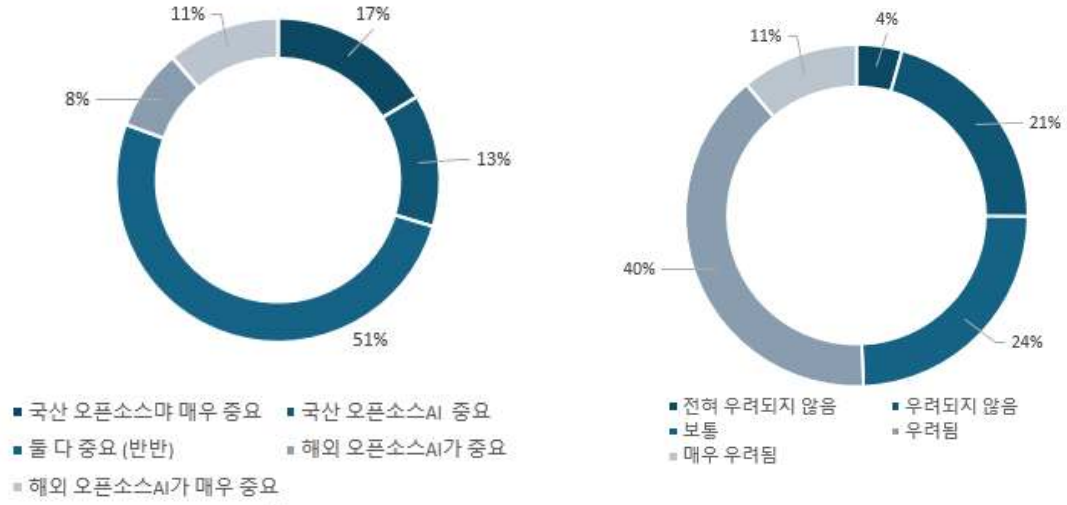
국산 오픈소스AI 중요성은 해외 오픈소스AI와의 중요성을 비교하기 위한 5개의 응답(국산 오픈소스AI가 매우 중요, 국산 오픈소스AI가 중요, 둘 다 중요, 해외 오픈소스AI가 중요, 해외 오픈소스AI가 매우 중요)로 구분하였다. 설문 결과, 가장 많은 선택을 받은 응답은 둘 다 중요로 189명(50.9%)가 응답하였다.

그리고 국산 오픈소스AI 중요(매우 중요 포함)하다는 응답은 110명으로 29.6%이었으며, 해외 오픈소스AI 중요(매우 중요 포함)하다는 응답은 72명으로 19.4%이었다. 이러한 응답 결과는 둘 다 중요하다고 볼 수 있는 응답이며 AI 개발자 입장에서는 우수한 성능과 비용 절감할 수 있는 좋은 오픈소스AI 기술이며 국산과 외산을 구분하지 않음을 알 수 있다.

특정 플랫폼과 기업에 대한 종속성 우려를 측정하기 위해 5점 리커트 척도(1=전혀 우려되지 않음, 2=우려되지 않음, 3=보통, 4=우려됨, 5=매우 우려됨)를 사용해 응답 결과를 수집하였다. 설문 결과에서 가장 많은 선택을 받은 응답은 우려됨으로 147명(39.6%)가 응답하였다. 그리고 우려되지 않음(전혀 우려되지 않음 포함)에 대한 응답은 총 93명으로 25.1%이었으며, 보통은 90명으로 24.3%이었다.

우려됨으로 응답한 총 응답자 수는 188명으로 50.7%로 전반적으로 플랫폼 및 특정 기업 종속에 대한 우려가 개발자들에게 있음을 확인하였다. 이는 기술 내재화 없이 단순 활용할 경우에 해당 기업이 추가적인 공개를 중단할 경우를 우려하는 것으로 판단되며, 자주적인 AI 역량 확보 필요성이 있다.

[그림 56] 국산 오픈소스AI 중요성과 특정 플랫폼/기업 종속성 우려



(a) 국산 오픈소스AI 중요성 (b) 특정 플랫폼/기업 종속성 우려

(자체 작성)

10) 오픈소스AI 전문가 양성 필요성 및 전문가의 핵심 역량

오픈소스AI 생태계가 빠르게 확산되면서 오픈소스AI 전문가의 필요성과 핵심 역량에 대한 관심이 증가하고 있다. 이에 따라 설문지 전문가 검토 과정에서 해당 설문들에 대한 요청을 반영하여 해당 설문 문항들을 추가하였다.

오픈소스AI 전문가 양성의 필요성을 측정하기 위해 설문 문항은 5점 리커트 척도(1=매우 중요함, 2=중요함, 3=보통, 4=중요하지 않음, 5=매우 중요하지 않음)을 사용해 응답 결과를 수집하였다. 설문 결과에서 가장 많은 응답은 매우 중요함으로 206명(55.5%)가 응답하였으며, 이어서 중요함 121명(32.6%)이 응답하였다. 따라서 전체 응답자의 88.1%가 오픈소스AI 전문가 양성 필요성을 제기하고 있다. 이러한 결과는 선행 설문 문항에서 인력 부족이 오픈소스AI 활용을 저해하는 2번째 요인으로 응답되었으며, 전문가 양성 및 발굴이 오픈소스AI 활용 촉진을 위한 2번째 요인으로 응답한 결과와 일관성이 있는 응답이다.

그리고, 오픈소스AI 전문가의 핵심 역량에 대한 설문은 7개의 응답들(모델 개발 및 최적화, 오픈소스 프레임워크 활용, 오픈소스 기여 및 협업, 데이터 처리 및 품질 관리, AI 윤리/보안/라이선스, 글로벌 커뮤니티 활동, AI 기반

제품/서비스 기획 및 관리)을 제안받아 구성하여 응답자가 최대 3개의 응답을 선택할 수 있도록 하였다. 설문 결과에서 가장 많은 응답을 받은 항목은 AI 모델 개발 역량으로 270명(72.8%)이 응답하였다. 이어서 오픈소스 프레임워크 활용 역량이 190명(51.2%)가 응답하였고, 데이터 처리 및 품질 관리, 데이터셋 구축 역량은 178명(48.0%), 오픈소스 기여 경험 및 협업 역량은 160명(43.1%), AI 윤리/보안/라이선스 등 규범적 이해는 110명(29.6%)로 응답되었다.

이러한 결과는 개발자 중심의 설문이어서 개발 역량 중심의 응답이 많았을 것으로 추정하며, AI 핵심 개발 역량은 AI 모델 개발, 오픈소스 프레임워크, 데이터 처리 및 관리 등과 같은 실무적인 AI 개발 역량에 대한 필요성이 높게 응답되었다. 그리고, 오픈소스 기여 및 글로벌 커뮤니티 활동에 대한 관심도가 낮아 오픈소스AI 활용에 대한 필요성이 더 높게 응답되었다.

[그림 57] 오픈소스AI 전문가 육성 필요성과 전문가의 핵심 역량

(a) 오픈소스AI 전문가 육성 필요성

(b) 오픈소스AI 전문가의 핵심 역량

(자체 작성)

제3절 요약 및 시사점

1. 요약

본 장은 국내외 오픈소스AI 생태계에 대한 실증적 분석을 진행하였다. 우선 EpochAI의 유명 AI 모델 데이터를 기반으로 글로벌 AI 생태계에서 오픈소스 모델의 영향력과 현황을 계량적으로 분석하였다. 그리고, 국내 오픈소스AI 인식 및 현황 파악을 위해 371명의 국내 AI 개발자를 대상으로 오픈소스AI 활용과 관련된 인식과 현황에 대한 설문 문항을 개발하여 조사를 진행하였다.

글로벌 생태계 분석을 위한 EpochAI의 유명 AI 모델 데이터는 1950년부터 2025년 6월 초사이에 발표된 AI 모델 중 4가지 조건(① 공인 벤치마크의 최첨단 개선, ② 인용 수 1000개 이상, ③ 기술 발전의 중요성, ④ 중요한 활용)중 하나 이상을 충족한 964개의 AI 모델 정보를 제공한다. OSI의 오픈소스AI 정의를 기반으로 주요 공개 항목(데이터, 모델 구조, 웨이트, 학습 & 추론 코드 등)을 중심으로 오픈소스 모델 269개를 분류하여 참여 기관, 참여 기관 국가, 조직 분류, 발표 일시, 유형, 활용 분야, 선정 기준 등을 분석하였다.

참여 기관에서 가장 많은 유명 모델 개발에 참여한 기관은 구글로 타 기관 대비 압도적이었으며 오픈소스 모델 공개는 메타(페이스북 포함시)가 45개로 1위를 차지하였다. 그리고, 기관 유형은 산업계가 참여한 AI 모델이 626개 이었으며, 오픈소스 모델은 208개로 기업들이 AI 기술 혁신을 주도하고 있었다.

연도별 현황을 보면 AI 모델은 2013년부터 지속적으로 증가하고 있었으며, 오픈소스 모델은 2015년에 처음 등장하여 2018년부터 지속 증가하는 추세이었다. 그리고, 2019년이후 유명 모델의 51.9%는 오픈소스 모델일 정도로 오픈소스 생태계 영향력이 최근 증가하였다. 활용 분야는 언어 모델링(1위)과 언어 모델링/생성(2위)를 차지하였다.

모델별 참여 국가 현황을 보면 미국이 AI 모델과 오픈소스 모델 모두 1위였고 이어서 중국이 2위였는데, 이러한 현황은 최근 AI 생태계에서 중국 기업들의 부상과 관련된 정보이다. 실제로 허깅페이스에서 중국산 모델들이 전세계적으로 주목을 받고 있으며 미국이 이에 대응하기 위한 움직임이 있

다.

데이터 정보 공개, 모델 접근성, 학습 코드 접근성, 추론 코드 접근성의 분석 결과를 보면 유명 모델의 약 절반인 48.3%(458개) 모델들이 학습 데이터 정보를 공개하고 있으며 학습 데이터 종류는 276개로 다양한 데이터들이 활용되고 있음을 알 수 있다. 이에 반해 모델 접근성, 학습 코드 접근성, 추론 코드 접근성을 공개한 모델은 각각 269개, 237개, 225개로 상대적으로 적음을 알 수 있다.

국내 오픈소스AI 인식 및 활용 현황 조사는 해당 설문 조사를 위해 적절한 조사 대상 선정과 설문 문항을 연계하여 기획되었으며, 모바일 온라인 설문 조사를 통해 총 371명의 응답자로부터 응답을 수집하였다. 주요 구성은 Part 1은 성능 기대, 노력 기대, 사회적 영향, 촉진 조건, 저항 요인, 활용 의사, 사용 행동, 기술 적합성, 환경 역동성, 제품 난이도에 대해 40개의 리커드 척도 문항으로 구성하였다. Part2는 오픈소스AI 관련 개발자들의 직접적인 인식과 상세 현황을 위한 설문으로 핵심 공개 항목, 중요성, 만족도 및 향후 예측, 국산 중요성, AI 도입 유형, 적용 분야 및 활용 수준, 모델 유형, 장단점, 저해 요인 및 촉진 조건, 플랫폼/기업 종속성, 전문가 필요성 및 핵심 역량에 관한 20개의 문항으로 구성하였다.

Part1의 오픈소스AI 활용 동기 측면에서 성능 기대에서 업무 생산성과 빠른 업무 수행에 상대적으로 많은 응답자들이 동의하였으며, 노력 기대에서 필요 정보 획득에 가장 많은 응답자들이 선택하였으며, 사회적 영향에서 의사결정자와 고위 경영진이 상대적 으로 많은 응답자들이 동의하였고, 촉진 조건에서 필요 지식이 가장 많이 응답되었다. 저해 요인으로는 저작권 침해에 가장 많은 응답자들이 동의하였다.

AI 개발자들의 오픈소스AI 활용 현황을 보면 활용 의사에서 보면 주변 추천, 자주 사용, 지속 사용, 방법 탐색 등 다양한 측면에서 활용을 확산할 의사를 가지고 있음을 확인하였다. 그리고 실질적 행동으로 기술 문서 탐색, 새로운 정보 획득, 학습 자료 활용 등을 통해 오픈소스AI 역량을 강화하려는 응답 비중이 50%를 넘었다.

Part2 부분에서 개발자들은 오픈소스AI의 핵심 공개 항목으로 모델, 데이터 정보, 학습 데이터, 기술 문서를 선택하였다. 이는 공개된 정보를 기반으로 활용과 성능 향상을 위한 핵심 정보들도 판단된다.

다양한 측면의 오픈소스AI 중요성에 대해 개인적인 업무상 중요성 외에 국가 AI 기술력 강화, 기업 AI 경쟁력 강화, AI 기술 확산 측면에 오픈소스AI 중요하다는 응답에 90% 안팎으로 응답할 정도로 중요하게 인식하고 있었다. 이러한 이유 중에 하나는 오

픈소스AI에 만족하는 개발자 비율이 54.4%로 과반이 넘었고, 향후 활용이 확대될 것으로 예상하는 개발자 비율이 89%이었다. 실제로 AI 원천 기술 도입 형태에서 오픈소스 AI를 활용하는 비율은 68.5%이었으며, 그 이유는 빠른 기술 확보, 개발 속도 향상, 개발 비용 절감 같은 R&D 효율성 측면이었다.

적용 분야와 활용 수준을 보면 사내 업무 및 서비스 적용과 기존 제품/서비스 개선이 각각 60.4%와 59.0%로 새로운 제품/서비스 개발에 활용하는 비율보다 높았으며, 활용 수준도 제품 적용 보다는 기술 탐색 및 연구개발 진행 중이 전체의 72.3%로 아직까지는 제품 적용 단계는 소수인 것으로 파악되었다.

그리고 오픈소스AI 기술의 장단점으로 빠른 기술 발전 속도, 라이선스 비용 절감, 및 기술 내재화 가능성에 많은 응답자들이 선택하였으며, 단점으로는 기술 완성도 부족, 책임 주체의 불명확성, 부족한 문서화 등이 선택되면서 기술 역량 확보가 오픈소스AI 확산의 주요 요소로 판단된다.

Part1의 오픈소스AI 저항 요인과 촉진 조건을 상세히 묻는 설문에서 주요 저해 요인으로는 조직의 오픈소스AI 전략(비전) 부재, 인력 부족, 자체 인프라 부족, 조직의 정책/거버넌스 부재 순으로 응답되며 개인적 오픈소스AI 활용 의사에 비해 환경적 요인이 부족한 것으로 판단된다. 그리고 주요 촉진 조건으로는 연구개발 인프라, 전문가 양성 및 발굴, 기업/기관 거버넌스 확립, 오픈소스AI 활용 기업 발굴 등이 응답되며 오픈소스AI 활용 기반 조성 및 인력 양성 필요성이 제기되었다.

기술 종속성 관점에서 국산 오픈소스AI의 중요성과 특정 플랫폼/기업 종속성에 대한 설문 결과를 보면 국산 오픈소스AI 기술이 중요하다고 응답한 비율이 해외 오픈소스AI가 중요하다고 응답한 비율보다 10.2% 높았으나, 둘 다 중요하다고 응답한 비율이 절반이 넘는 50.9%임을 고려하면 개발자들은 국산 혹은 외산 구분 보다는 우수한 오픈소스AI에 관심이 많다고 추정할 수 있다. 다만, 특정 플랫폼 및 기업 종속성에 대한 우려에 대한 응답은 50.7%로 이에 대한 우려가 있음을 확인하였다.

마지막으로 오픈소스AI 확산의 주요 저해 요소인 인력(개발자, 기여자) 부족이자 주요 촉진 조건인 전문가 양성을 위한 오픈소스AI 전문가 육성 필요성에 중요하다고 응답한 비율은 88.1%로 많은 개발자들이 동의하고 있었다. 그리고, 전문가의 주요 역량으로 AI 모델 개발, 오픈소스 프레임워크 활용, 데이터 처리 및 품질 관리와 데이터셋 구축 역량, 오픈소스 기여 및 협업 역량 순이었다.

2. 시사점

1) 빠르게 성장하는 오픈소스AI 생태계

오픈소스AI 기술은 AI 기술 혁신과 확산의 원동력으로 이러한 현상은 AI 기술의 핵심인 모델 분야에서도 확인되고 있다. EpochAI가 선정한 유명 AI 모델에서 오픈소스 모델 비중은 28.4%에 불과하지만 2019년 이후로 한정할 경우에 그 비율은 51.8%로 크게 증가한다. 이러한 오픈소스AI 생태계 성장 요인에는 국가로는 미국과 중국이 있고 기업으로는 구글, 메타, 알리바바 등이 있다.

실제로 유명 AI 모델의 미국 기관 참여율 66.9%로 압도적이며, 이어서 중국 기관 참여율은 14%로 영국 캐나다와 큰 차이가 없다. 오픈소스 모델의 경우 미국 기관 참여율은 69.1%로 다소 높아지고 중국 기관 참여율도 21.2%로 다소 증가한다. 유명 AI 모델 개발의 기업 참여 비율은 66%이며, 오픈소스 모델의 기업 참여 비율은 77.3%이다.

그리고, 유명 AI 모델의 유형은 최근 각광받고 있는 생성형 AI 분야의 언어 모델, 비전 모델, 멀티모달 모델, 이미지 생성 모델이며, 활용 분야도 이와 유사하게 언어 모델링, 언어 모델링/생성, 이미지 분류, 질의 응답, 번역 분야이다. 이는 유명 AI 모델과 오픈소스 모델이 모두 유사하며 오픈소스 모델이 AI 생태계와 많은 접점을 가진다고 해석할 수 있다.

2) 기업/국가 경쟁력 강화를 위해 중요한 오픈소스AI

오픈소스AI 기술이 단순히 양적인 성장 뿐만 아니라 질적으로도 중요해지고 있다고 판단된다. 1차적으로 국내 오픈소스AI 설문 결과에서 국가 AI 정책 측면의 4가지 요소(업무 상 중요성, 국가 AI 기술력 강화, AI 기술 확산) 모두에서 중요성이 최소 87% 이상이었다. 특히 기업 AI 경쟁력 강화, 국가 AI 기술력 강화, AI 기술 확산 측면의 중요성을 동의하는 비율은 90%를 넘어서며 개인 업무 상 중요성 보다 중요성을 크게 인식되고 있다.

이는 AI의 중요성(산업적 영향력, 국가 미래 경쟁력 등) 커져감에 따라 개발자들이 AI 기술을 확보할 수 있는 현실적 방안으로 오픈소스AI 기술을 중요하게 생각하고 있기 때문으로 판단된다. 실제로 오픈소스AI 성능에 만족을 느끼는 개발자 비율이 과반을 넘어서며 향후에 오픈소스AI 활용이 증가

할 것으로 예측하는 개발자 비율이 거의 90%에 달하고 있다. 그리고 AI 원천 기술 도입 형태로 오픈소스AI를 활용하는 비율이 68.5%일 정도로 국내에서 오픈소스AI 활용이 보편화되고 있다.

또한 오픈소스AI의 주요 장점으로 빠른 기술 발전 속도, 공개된 기술의 내재화 가능성, 기술 투명성 및 검증 가능성 등이 주요 장점으로 응답되면서 오픈소스AI 기술에 대한 긍정적 인식이 개발자 사이에 있기 때문이다.

3) 기업 인식 개선 필요

글로벌 오픈소스AI 생태계 현황을 보면 미국과 중국 기업들을 중심으로 적극적 연구개발과 결과물 공개가 추진되며 오픈소스AI 생태계가 지속적으로 성장하고 있다. 실제로 EpochAI의 유명 AI 모델의 66%는 기업이 참여하여 개발한 모델이며, 특히 오픈소스 모델의 경우에는 기업 참여율은 77.3%로 더 높은 현실이다. 그리고 오픈소스 모델의 국가별 순위를 보면 우리나라는 8위 이기 때문에 더 많은 투자가 필요하다고 볼 수 있다.

그리고, 국내 인식과 현황 조사의 Part1을 보면 오픈소스AI 활용에 영향을 주는 사회적 요인을 보면 의사결정자(상사)와 고위 경영진이 1위와 2위를 차지하며 회사 내부 의사 결정 구조의 중요함을 제시하고 있다. 그리고, 오픈소스AI 활용의 주요 저해 요인으로 조직의 오픈소스AI 전략(비전) 부재, 인력 부족, 자체 인프라 부족 등과 같은 기업 거버넌스 요소와 밀접한 요인들로 파악되고 있으며, 주요 촉진 조건으로 연구개발 인프라 제공, 기업/기관의 거버넌스 체계 구축 같은 조직적 측면의 요인들이 거론되고 있다.

국내외 현황을 보면 오픈소스AI가 회사 업무 성과 향상에 긍정적인 면을 미치고 있음에도 개발자들은 조직적 지원이 부족하다고 인식하고 있다. 따라서, 오픈소스AI 활성화를 위해서는 기업의 오픈소스AI 인식 개선과 지원 체계 마련이 필요하다고 판단된다.

4) 오픈소스AI 활용을 위한 전제조건 : 기술 역량 확보

오픈소스AI는 기술을 공개하여 다양한 장점들(기술 발전 속도, 기술 내재화 가능성, 업무 효율 향상 등)들이 있음에도 이를 제대로 활용하기 위해서는 충분한 기술적 역량을 필요로 한다. 실제 설문 결과의 주요 단점으로 주요 단점으로 기술적 완성도 부족(45.3%), 책임주체 불명확(37.5%), 부족한

문서화, 연구개발 난이도(부담 증가, 29.9%), 보안 및 운영 리스크(29.4%) 등 다양한 기술적 이슈들이 제기되었다.

결국은 오픈소스AI를 적극적이고 효율적인 활용을 위해서는 오픈소스AI의 단점을 해결하거나 극복할 수 있는 기술적 역량이 중요하다. 이러한 기술적 장벽들을 비즈니스 관점에서 보면 오픈소스AI 기업의 한계이자 전략적 선택이라고 볼 수 있다.

왜냐하면 기업은 이윤 창출을 위한 조직인데, 모든 기술을 공개할 경우 수익화가 어려울 수 있기 때문에 인위적(전략적)인 기술적 장벽을 통해 기술력이 부족한 기업들을 대상으로 비즈니스 기회를 창출할 수 있다. 따라서 오픈소스 모델의 기술적 장벽을 제거할 수 있는 기업은 오픈소스 모델의 비즈니스 기회를 얻을 수 있게 된다.

5) 오픈소스AI 인재 양성 필요

오픈소스AI를 적극적이고 효율적 활용을 위한 기술적 역량의 근간은 오픈소스AI 전문가이다. 실제로 국내 설문 조사에서 오픈소스AI의 주요 저해 요인으로 인력(핵심 개발자, 기여자) 부족이 응답되었으며, 주요 촉진 조건으로 전문가 양성(기술 교육) 및 발굴이 응답되었다. 이러한 응답은 국내에 오픈소스AI 전문가가 부족하다는 것을 의미한다.

그리고, 향후 오픈소스AI 활용이 증가할 것으로 예측하는 개발자의 비중 89%인 것을 감안하면 우수한 기술 역량을 갖춘 오픈소스AI 전문가 수요는 더욱 증가할 것으로 예상된다. 따라서, 오픈소스AI 인재 양성은 국내 오픈소스AI 활성화를 위한 중요 전제 조건이라고 할 수 있다.

6) 자주적 AI 역량 확보가 필요

절반 이상의 개발자들이 국산 오픈소스AI와 해외 오픈소스AI 모두 중요하다고 응답하고 해외 오픈소스AI가 중요하다고 응답한 비중도 20%에 가까울 정도로 국산 오픈소스AI와 해외 오픈소스AI에 대한 뚜렷한 선호의 차이가 보이지 않았다. 하지만, 특정 플랫폼과 기업 종속성에 대한 우려에서는 우려된다고 밝힌 응답자 비중 45%일 정도로 우려되지 않는다고 밝힌 개발자(15%)보다 훨씬 높았다.

이러한 응답 배경에는 현재 공개된 오픈소스AI 기술에 대한 우려는 크지

않지만, 향후 특정 플랫폼/기업의 정책 변화에 따른 잠재적 우려가 있는 것으로 판단된다. 실제로 과거 OpenAI는 GPT-1/2 기술은 공개하였지만, chatGPT 서비스 출시 이후 비공개로 전환하였으며, 메타도 2025년도에 기술 비공개 전환을 논의한다는 이야기가 있었다. 이러한 사례들 때문에 개발자들이 특정 기업의 정책 변화(비공개 전환)에 대한 우려가 있는 것으로 판단된다.

따라서, 기업 종속성 우려를 완화하기 위한 자주적인 AI 역량 확보가 필요하며, 이를 위한 수단으로 오픈소스AI 기술 활용 및 선진 오픈소스AI 기술 내재화를 추진할 필요가 있다.

제5장 결론: 정책적 시사점 및 정책 제언

AI 시대에 오픈소스 생태계의 중요성이 더욱 커지고 있다. 실제로 글로벌 기업의 89%가 AI 개발 과정에 오픈소스 기술을 활용하고 있을 정도로 오픈소스 기술은 AI의 기술적 기반을 제공하고 있다. 텐서플로우와 파이토치 같은 오픈소스 AI프레임워크가 머신러닝 시대를 개막하고 모델 대형화라는 기술 혁신을 넘어 생성형AI 산업을 태동시키는 산업 혁신의 원동력이 되었다.

최근 AI 선도기업들은 영향력 확대를 위해 적극적으로 오픈소스 모델을 공개하며 AI 기술 내재화가 가능한 원천 기술들을 적극적으로 공급하고 있다. 오픈소스 생태계는 AI 시대의 기술적 기반을 제공하며 AI 기술 혁신과 산업 혁신의 원동력이 되고 있다. 실제로 국내 AI 개발자의 절대 다수가 오픈소스AI가 기업 AI 경쟁력 강화(93.5%), AI 기술 확산(92.7%), 국가 AI 기술 경쟁력 강화(91.1%), AI 연구개발 업무 측면(87.3%)에서 모두 중요하다고 응답하였다.

따라서, 기업 AI 경쟁력 강화, 국가 AI 기술력 강화 등을 위한 오픈소스AI 역할이 중요해지면서 AI 3강 도약을 위한 오픈소스AI 정책 필요성이 더욱 커지고 있다. 따라서 본 연구 내용을 기반으로 다음과 같은 정책적 시사점과 정책을 제언하고자 한다.

1) 선진 기술의 빠른 수용을 위한 오픈소스AI 활용 확산 지원

오픈소스AI 기술은 AI 기술 혁신과 산업 혁신의 원동력이 되고 있다. 오픈소스AI 기술은 새로운 AI 기술 패러다임을 제시하며 AI 기술 혁신을 선도하고 있기 때문이다. 실제 트랜스포머, BERT 등은 AI 모델의 핵심 구조를 제안하였고 딥시크는 고효율 MoE 기술을 제시하며 모델 경제성 패러다임을 선도하였다. 그리고 상용AI 모델과 유사 성능의 오픈소스 모델은 낮은 비용으로 AI 서비스 환경을 구축 가능하게 하여 AI산업 성장을 자극하고 있다.

따라서, 선진 오픈소스AI 기술을 빠르게 수용하기 위한 오픈소스AI 활용

확산 정책의 필요성이 제기되고 있다. 선진 오픈소스AI 기술은 공개된 기술로 기술 접근성이 우수하며 상용AI 서비스와 유사한 성능을 제공하고 있다. 이를 효과적으로 활용한다면 AI 제품・서비스 품질 향상과 AI 개발 속도 향상을 기대할 수 있고 공개 검증을 거친 오픈소스AI 기술로 제품・서비스 신뢰성 향상에 도움이 된다. 또한 상용AI 솔루션과 달리 자체 구축을 통한 외부 의존도 완화 및 비용 절감(토큰 비용 등)을 기대할 수 있다.

그러나, 오픈소스AI 기술은 누구나 활용 가능하지만, 실제로는 기술력이 부족한 개발자나 기업은 활용하기 어렵다는 단점이 있다. 그 이유는 상용 기술이 아니기 때문에 맞춤형 기술 지원이 불가능하여 최적화, 편의성, AI 신뢰성 등을 자체적으로 해결해야 하기 때문이다. 이 과정은 AI 제품 및 서비스의 완성도를 높이는 실질적인 수익화 개발 단계로 상용AI 기술을 활용한다면 비용 지불을 통한 기술적 애로사항을 쉽게 해결할 수 있다. 하지만, 오픈소스AI 기술을 활용할 경우에는 비용 절감 대신 기업 스스로 문제를 해결하고 하는 어려움이 있다.

따라서, 오픈소스AI 활용 확산은 SW산업 측면에서 대외 의존도 완화라는 긍정적 효과를 기대할 수 있는 가장 현실적인 방안이다. 공개적으로 검증된 우수한 오픈소스AI을 활용하는 기업들이 증가할수록 기업들의 AI 기술력이 향상되고 대외 의존도를 완화할 수 있기 때문이다. 대외 의존도 완화는 개별 기업 수익 향상을 통한 SW산업 경쟁력 향상과 자체적인 AI 사업화를 통한 SW 인력 고용 확대 효과까지 기대할 수 있다.

그러므로, 오픈소스AI 활용 확산은 AI 3강 도약을 위한 저변 확대 및 SW산업 내실화를 위해 매우 중요한 과제이며 오픈소스AI 활용 촉진과 활용 어려움 완화를 위한 정책적 지원이 필요하다. 이러한 정책 추진에 있어 국가 생태계 전반에 AI 경쟁력을 제고하기 위한 생태계 관점의 정책이 필요하다. 즉, AI 기술력 강화를 위한 선진 오픈소스AI 기술 내재화 추진, 기업 오픈소스AI 인식 개선, 오픈소스AI 전문가 양성, 오픈소스AI 전문가 컨설팅 제공, 오픈소스AI 전문기업 육성 등의 종합적인 정책 추진이 필요하다.

2) AI 기술력 강화를 위한 선진 오픈소스AI 기술 내재화 추진

글로벌 기업의 89%가 AI 개발 과정에 오픈소스 기술을 활용하고 있을 정도로 오픈소스 기술은 AI의 기술적 기반을 제공하고 있기 때문에 오픈소스AI 기술은 AI 개발을 위한 필수 기술로 볼 수 있다. 하지만 과거 우리나

라는 오픈소스를 SW 라이선스 비용 절감 수단으로 인식하여 개발자 개인의 단순 활용 수준에 머무르며 성능 개선 및 최적화에 어려움이 있었다.

AI 분야도 SW와 같이 단순 활용을 벗어나지 못할 경우에는 재학습 및 성능 최적화에 한계를 보이거나 외부 의존성이 강화될 수 밖에 없다. 따라서, 핵심 기술을 이해, 검증, 개선, 확장 할 수 있는 완전한 기술 내재화가 필요하다. 특히 일부 기업들은 오픈소스 모델 공개를 통해 영향력 확대를 추진하다가 개발자들의 의존성이 강화되면 이를 비공개 상용 서비스로 전환하는 미끼 상품 전략을 활용하기 때문에 이러한 기업의 기술 마케팅 전략을 벗어나기 위해서는 선진 오픈소스AI 기술 내재화가 필요하다.

선진 오픈소스AI 기술 내재화를 위해서는 전략적 목적(원천 기술 확보, 호환 기술 개발, 성능 향상, 제품/서비스 개발 등)을 명확히 하여 시장 지배적인 오픈소스 기술 기반의 R&D를 추진할 필요가 있다. 그리고, 단순 포크(fork)가 아닌 구조 분석・성능 비교・적용 가능성 검토를 고려한 심층 기술 분석 체계를 기반으로 해야 한다. 특히 최근에는 모델 성능과 함께 비용 효율성이 강조되기 때문에 학습 최적화, 추론 최적화, 경량화 등을 추진하여 기술 종속성을 벗어나 자주적인 수정, 확장, 대체가 가능한 수준의 역량 확보가 필요하다.

3) 기업의 오픈소스AI 인식 개선 필요

국내 개발자의 설문 조사에서 절반 이상의 개발자들이 오픈소스AI에 만족을 하고 있었으며, 대부분의 개발자들은 향후 오픈소스AI의 활용이 확대될 것으로 예상하고 있었다. 그러나, 오픈소스AI 활용 확산의 주요 저해 요인으로 조직의 오픈소스AI 전략(비전) 부재와 조적의 정책/거버넌스 부재가 응답되었다. 이 결과는 개발자들이 오픈소스AI 활용이 업무 효율성과 성과에 유리하다고 인식하는데 반해 기업 내부 체계가 개발자들을 체계적으로 지원하고 있지 못함을 의미한다.

그리고, 사회적 영향 분석에서 오픈소스AI 활용에 가장 큰 영향을 미치는 주체가 회사의 의사 결정자에 해당하는 상사와 고위 경영진으로 응답되면서 기업의 의사 결정 구조가 기업 내부의 오픈소스AI 활용 확산에 크게 영향을 주고 있음을 알 수 있다. 또한, 이 설문 결과는 오픈소스AI 활용 확산의 주요 저해 요인들인 오픈소스AI 전략 및 정책/거버넌스 부재와 관련성이

있다. 그 이유는 기업 의사결정자들이 오픈소스AI에 대한 인식이 긍정적이라면 오픈소스AI 전략 및 정책/거버넌스에 많은 관심을 가질 수 있는데 반해 기업 의사결정자들이 오픈소스AI 활용 확산의 중요성에 인지하지 못하였다면 오픈소스AI 전략 및 정책/거버넌스 체계가 부재할 가능성이 높기 때문이다.

따라서, 기업 오픈소스AI 활용 확산을 위해서는 기업 경영진과 의사결정자들의 오픈소스AI에 대한 인식 개선이 필요하다. 기업 (경영진과 의사결정자의) 인식 개선을 위해서는 1차적으로 오픈소스AI의 전략적 가치와 사업화 방안에 대한 교육 제공이 필요하고, 오픈소스AI 거버넌스 컨설팅 및 구축 지원이 필요하다.

특히 오픈소스AI 거버넌스는 오픈소스AI 활용에 따른 위험 예방(저작권 침해, 개인정보 침해, 보안 등)을 위한 체계이기 때문에 개발자들의 오픈소스AI 활용의 부담을 완화해줄 수 있다. 또한 오픈소스 사업화의 모범 사례 발굴을 위한 오픈소스AI 전문 기업 육성을 추진할 필요가 있다. 이는 교육을 통한 점진적인 인식 개선 보다 기업 활동에 도움이 된다는 실질 사례를 발굴함으로써 기업의 오픈소스AI 인식 개선에 큰 효과를 볼 수 있을 것으로 판단된다.

4) 오픈소스AI 전문가 양성 및 교육 확대 필요

국내 인식 및 현황 설문 조사에서 오픈소스AI 활용의 주요 저해 요인으로 인력(핵심 개발자, 기여자) 부족과 개인 기술 역량의 한계가 응답되었으며, 오픈소스AI 활용의 주요 촉진 조건으로 전문가 양성(기술 교육) 및 발굴이 2번째로 응답되었다.

이는 많은 개발자들이 오픈소스AI 활용에 어려움을 겪고 있기 때문이다. 해당 설문 조사에서 오픈소스AI의 주요 단점으로 기술적 완성도 부족, 활용 시 책임 주체 불명확, 부족한 문서화로 인한 접근성 한계 등이 제기되었으며, 전문가 양성 필요성에 89%의 개발자들이 중요하다고 응답할 정도로 오픈소스AI 전문가 양성 필요성이 매우 높게 제기되었다.

따라서, 오픈소스AI 전문가 발굴 및 양성을 위한 정책 추진이 필요하다. 설문 조사 결과에서 전문가의 핵심 역량으로 AI 모델 개발 및 최적화, 오픈소스 프레임워크 활용 역량, 데이터 처리 및 품질 관리, 데이터 셋 구축 역량이 주요 역량으로 응답되어 AI 모델 개발 및 최적화를 위한 역량들이 매

우 중요하다고 판단된다. 따라서 이러한 역량을 갖춘 전문가 양성을 위해 문제 해결형 교육(코드 분석 및 개선, 파생 모델 개발 등) 확대, 산업 수요 연계형 오픈소스AI 인재 양성이 필요하며, 해커톤・개발자 대회 등을 개최를 통한 오픈소스AI 전문가 발굴도 필요하다.

5) 오픈소스AI 전문 컨설팅 제공 필요

오픈소스AI 기술을 포함한 AI 기술은 기술적 난이도 높다고 인식되어 있다. 특히 오픈소스AI 기술은 자발적 공개에 따른 기술적 완성도가 높지 않기 때문에 이로 인한 다양한 문제들을 내포하고 있을 수 있다. 국내 인식 및 현황 설문 조사에서 오픈소스AI의 주요 단점으로 기술적 완성도 부족, 책임 주체 불명확(편향, 윤리), 부족한 문서화로 인한 접근성 한계, 복잡한 라이선스 정책 및 저작권 침해, 보안 및 운영 리스크, 연구개발 난이도 상승(자체 인프라 구축)과 같은 다양한 기술적 또는 시스템 운영 관점의 애로사항들이 응답되었다.

이러한 설문 결과는 개발자들이 오픈소스AI 활용에 다양한 어려움을 겪고 있는 것으로 판단되며, 이러한 다양한 애로사항들은 교육만으로 해결하기에는 복잡하기 때문에 전문 컨설팅을 통한 애로사항 해결이 필요하다. 따라서 오픈소스AI 전문 컨설팅을 지원하여 기술적, 운영적 애로사항 해결을 지원해줄 필요가 있다. 주요 컨설팅 항목으로는 기술적 이슈와 거버넌스적 이슈 모두를 포괄해서 제공하는 것이 오픈소스AI 활용 확산이 도움이 될 것이다.

6) 오픈소스AI 전문기업 육성

글로벌 오픈소스AI 생태계는 AI 선도 기업들이 주도하며 성장하고 있다. 대표적 기업으로는 메타, 구글, 알리바바, 바이두 같은 빅테크 기업들도 있지만, 오픈소스 전략을 통해 글로벌 인지도를 높인 미스트랄AI, 딥시크, OpenAI 같은 전문 기업들도 있다. 이는 AI 산업 분야의 전망이 유망하기 때문에 해당 전문 기업들이 적극적 기술 개발을 통해 성장하고 있기 때문이다. 실제로 EpochAI의 유명 AI 모델 분석에서도 유명 오픈소스 모델의 다수는 기업들이 참여하여 개발된 것으로 AI 기업의 오픈소스 전략은 중요해지고 있다.

국내에서도 우수한 오픈소스AI 전문기업을 육성을 통한 AI 산업 활성화와 국가 AI

기업 경쟁력을 제고할 필요가 있다. 오픈소스AI 기업의 장점은 개방형 기술 검증을 통해 해당 기업의 기술력이 우수한지를 직접 검증할 수 있기 때문에 기술 난이도 높은 AI 기업 특성상 보다 투명하게 대상 기업을 선정하고 지원할 수 있는 장점이 있다. 또한 스타트업 입장에서는 자체 기술력을 오픈소스 전략을 통해 입증한다면 많은 AI 스타트업 중에서 국내외 인지도를 빠르게 향상시킬 수 있는 장점도 있다. 실제로 국내에서 업스테이지가 오픈소스 전략(솔라 모델의 글로벌 리더보드 1위 기록)으로 인지도를 향상시키며 문서 AI 및 온디바이스 AI 분야에서 경쟁력을 인정 받고 있다.

이와 같이 오픈소스AI 기업 육성은 AI 산업 활성화와 투명하게 검증된 기업 지원이라는 장점을 가지고 있기 때문에 적극 추진할 필요가 있다. 지원 방식으로는 경쟁력 있는 모델을 가진 기업 지원, 해외 유명 오픈소스AI 기술 기반 기업 등 여러 대상 기업들을 대상으로 선정할 수 있다.

참 고 문 헌

[국내 문헌]

네이버 기술 블로그, HyperCLOVA X THINK: From seeds to forest, 2025.

네이버 기술 블로그, Introducing HyperCLOVA X SEED, a commercial open-source, 2025.

네이버, Naver Unveils “HyperCLOVA X THINK,” a Reasoning Model with Enhanced AI Agent Capabilities, 2025.

네이버 클라우드 블로그, HyperCLOVA X Use Cases: Legal Q&A, 2025.

업스테이지. Solar: Apache 2.0 open-source strategy and B2B success. 2025.

업스테이지, Solar Pro 2: Fluent. Reasoning. Frontier, 2025.

업스테이지 블로그, Solar Pro 2 Preview: Small. Powerful. Now with reasoning, 2025.

업스테이지, Solar Pro 2 Breaks Into Global Frontier AI, 2025.

연합뉴스, 네이버 forecast to post record annual revenue, AI Profitability Improving, 2025.

엔씨소프트 공식 뉴스룸, NCSOFT Unveils 'Llama-VARCO LLM', 2024.

엔씨소프트, “Establishment of 'NC AI' Spinoff for Commercialization“, 2024.

엔씨 Research, “VARCO LLM License (CC BY-NC 4.0) & Technical Report“, 2025.

엔씨 Research 기술 블로그, VARCO LLM 2.0 Details, 2024.

조선비즈, LG unveils EXAONE ecosystem and hybrid AI at AI Talk Concert 2025, 2025.

조선비즈, 네이버 invests over 1 trillion won in R&D in H1, Focusing on AI, 2025.

중앙일보, LG AI Research launches Korea's first open-source AI model Exaone 3.0. 2024.

중앙일보, Naver unveils homegrown AI model HyperClova X Think, 2025.

한국무역협회, IMF, 올해 세계성장률 3.2%로 0.2%P↑... “무역합의로 관세영향 ↓”, 2025.10.15.

Korea Economic Daily, NCSOFT launches evaluation model to verify performance, 2024.

Korea Bizwire, Naver Unveils Major AI Upgrades with HyperCLOVA X for Search, Shopping and Maps, 2025.
Korea Herald, LG AI Research Showcases 'EXAONE Ecosystem', 2025.
Korea Herald, LG, SK Telecom, Naver selected for Korea's sovereign AI push, 2025.
LG AI Research 공식 유투브, Meet ChatEXAONE: LG' s Enterprise AI Agent, 2025.
LG AI Research 기술 보고서, EXAONE 3.0 7.8B Instruction Tuned Model Performance, 2024.
LG AI Research, EXAONE 4.0: Unified Large Language Models, 2025.
LG AI Research GitHub, EXAONE 3.0 License & Repository, 2024.
SK텔레콤 공식 언론 보도, SK Telecom Unveils Proprietary Standard Large Language Model A.X 4.0, 2025.
SK텔레콤 뉴스룸, SK Telecom Unveils Plans for 'AI Infrastructure Superhighway', 2025.

[해외 문헌]

AIbase News, Baidu Open Sources the WENXIN Large Model 4.5 Series, 2025.
AI Competence, Mistral AI: Europe' s Bold Move For AI Sovereignty, 2025.
AInvest, Baidu's AI Ecosystem and ERNIE X1.1: A Strategic Catalyst for Long-Term Growth, 2025.
AInvest, DeepSeek Blows Up Meta's AI Strategy: A Paradigm Shift in the AI Race, 2025.
Alibaba Cloud, Qwen LLM Tops 90,000 Enterprise Clients, 2024.
Alibaba Cloud (Alizila), Alibaba Introduces Qwen3, Setting New Benchmark, 2025.
Alibaba Cloud Blog, Alibaba Launches Qwen App to Boost its Consumer AI Efforts, 2025.
AIFundingTracker, Mistral AI Revenue Growth: From Zero to $100M+, 2025,
Aivancity, VivaTech 2025: Mistral AI unveils sovereign HPC infrastructure, 2025.
Asian Intelligence, Solar Pro 2: South Korea's Frontier LLM, 2025.
Baidu ERNIE Blog, ERNIE 4.5 Model Family: Open Sourcing MoE Models, 2025.
BeBeez, Europe's AI bet: Paris-based Mistral wins €1.7 billion, doubling valuation to €11.7 billion, 2025.

Bloomberg, Amazon Joins AMD to Back South Korean AI Startup Upstage, 2025.
Bloomberg, Alibaba Plans to Spend $53 Billion on AI Infrastructure, 2025.
BNP Paribas, BNP Paribas and Mistral AI sign a partnership agreement, 2024.
BrainIllustrate, Mistral AI: The Open-Core Challenger Forging a New Path, 2025.
Business Standard, How much of Silicon Valley's AI boom is powered by China's models, 2025.
Captide, Meta Q4 2024 Earnings Analysis, 2025.
China Daily, WeChat embracing DeepSeek for tech leap (AI Search integration), 2025.
CNBC, Alibaba shares rise as AI drives 34% cloud sales jump, 2025.
CNBC, AI firm Mistral valued at $14 billion as chip giant ASML takes major stake, 2025.
CNBC, China's biggest public AI drop since DeepSeek, Baidu's Ernie, is about to hit the market, 2025.
CNN, A shocking Chinese AI advancement called DeepSeek is sending US stocks plunging, 2025.
CNBC, Nvidia stock plummets, loses record $589 billion as DeepSeek prompts questions over AI spending, 2025.
CNBC, OpenAI hits $10 billion in annualized revenue, 2025.
ConnectCX, Alibaba Cloud's Qwen: Powering China's AI-Driven Industrial, 2025.
Complete AI Training, Alibaba deploys Qwen AI across Taobao and Tmall for double-digit lifts, 2025.
Creative Strategies, DeepSeek MoE & V2: Commoditizing Large Models, 2025.
Database Mart, RTX 4090 LLM Inference Benchmark, 2025.
DataCamp, Qwen 2.5-Max: Features, DeepSeek V3 Comparison & More, 2025.
DB Engines, Popularity of open source DBMS versus commercial DBMS, 2025.11.20.
DeepSeek-AI, DeepSeek-V3 Technical Report, 2024.
EarningsIQ, Baidu Q1 2025: AI Cloud Jumps 42%, Reshaping Core Revenue Mix, 2025.
EpochAI, Compute trends across three eras of machine learning, 2022.05.02.
Forbes, OpenAI Launches GPT Store: Where Creators Can Share—And Possibly Make Money, 2024.
HelloT (IT News), SK Telecom unveils A.X 4.0: Higher token efficiency than

GPT-4o, 2025.
Hugging Face Blog, Finally, a Replacement for BERT: Introducing ModernBERT, 2025.
Hugging Face, NCSOFT Model Download Stats, 2025.
Hugging Face, Upstage SOLAR 10.7B Apache 2.0 License & Performance, 2025.
Hugging Face, SK Telecom Organization Profile & Download Stats, 2025.
Hugging Face, Model statistics: Mistral ranks 2nd in downloads after Meta, 2025.
Hugging Face, EleutherAI Pythia Suite, 2025.
Jina AI, What should we learn from ModernBERT?, 2025.
Galileo, DeepSeek R1 vs OpenAI o1 Comparison, 2025.
Gasgoo, Baidu's Apollo Go fast-tracks global push with Uber, Lyft alliances, 2025.
GitHub, Octoverse 2025, 2025.10.28.
Gitnux, Report 2025 - Server Statistics, 2025.04.29.
GNU Operating Systems, What is Free Software?.
Google Cloud, Gen AI Unicorns and Fortune 500 Adoption, 2025.
Google Cloud Press Release, LG AI Research Taps Google Cloud to Develop EXAONE 3.0 and ChatEXAONE, 2024.
Google Developers Blog, Gemma explained: What's new in Gemma 3, 2025.
Google DeepMind, Gemma 3 Technical Report, 2025.
GPT-OSS Documentation, Getting Started Guide: Function Calling, 2025.
Groq, Llama 4 Live Today on Groq — Build Fast at the Lowest Cost, 2025.
Kaitchup, OpenAI GPT-OSS: Native 4-Bit MoE Models, 2025.
Linux Foudattion, Model Openness Framework, 2025.
Linux Foundtaion & Meta, The Economic and Workforece Impact of Open Source AI, 2025.05.
Linux Foundtaion, The State of Soveriegn AI, 2025.08.
LlamaIMoel.com, Llama AI for Commercial Use: License & Enterprise Impact, 2025.
LMSYS Org, Chatbot Arena Leaderboard: Coding & Hard Prompts, 2025.
Narratize, Renault reduces design time by 50% with AI, 2024.
NDTV World, Alibaba Qwen2.5 Max beats rivals DeepSeek & GPT-4o, 2025.
Network World, Alibaba is developing an AI inference chip amid US export curbs, 2025.

Maeil Business Newspaper, NC AI announced VARCO 3D as core infrastructure for digital twin, 2025.
Maginative, Meta's Llama Hits 1B Downloads, 2025.
MarkTechPost, Gemma 3 License Guide (Custom License, Not Apache 2.0), 2025.
Matrix BCG, Competitive Landscape of Naver: Market Share & Financials, 2025.
Meta, Meta Connect 2025 & Our Inaugural LlamaCon (September 17–18), 2025.
Meta AI, Introducing Meta Llama 3: The most capable openly available LLM to date, 2024.
Meta AI Blog, With 10x growth since 2023, Llama is the leading engine of AI innovation, 2024.
Meta Investor Relations, Meta Reports Fourth Quarter and Full Year 2024 Results, 2024.
Miracuves, DeepSeek Revenue Model 2025 ($1.1B Revenue Forecast), 2025.
Moor Insights & Strategy, Meta Llama's Enterprise AI Value, 2025.
OECD, AI Principles overview.
OpenAI, Introducing gpt-oss: Pushing the frontier of open-weight reasoning models, 2025.
OpenAI, Morgan Stanley uses AI evals to shape the future of financial services, 2024.
OpenTools, Meta's Llama AI Hits 1.2 Billion Downloads: A New Milestone in Open-Source AI, 2025.
Orange, Orange and Mistral AI Forge Strategic AI Partnership, 2025.
Open Souce Initative, Open Source Definition,
Open Souce Initative, The Open Source AI Definition 1.0
Precedence Research, Artificial Intelligence(AI) Market Size, Share and Trends 2025 to 2034. 2025.09.29.
PR Newswire, Baidu Unveils ERNIE 5.0 and AI Applications at Baidu World 2025, 2025.
PR Newswire, Baidu Launches ERNIE 4.5 Turbo with competitive pricing, 2025.
Quantum Zeitgeist, Llama AI Model Sees Widespread Adoption Across Industries (Accenture, AT&T cases), 2025.
RD World, DeepSeek-R1 RL model: 95% cost cut vs. OpenAI's o1 ($5.6M training

cost), 2025.
Reddit, Deepseek R1 just became the most liked model ever on Hugging Face, 2025.
ReelMind AI, Gemini Nano On Device: AI's Mobile Technology, 2025.
Reuters, South Korean ministries block DeepSeek on security concerns, 2025.
Rohan Paul, Google released Gemma 3: 128k Long-Context Window, 2025.
Ryan O' Connor, PyTorch vs Tensorflow in 2023, Speech&Text, 2021.12.14.
Scribd, Mistral AI La Plateforme & API Usage Statistics, 2025.
SCMP, Alibaba holds wide lead over rivals (Baidu 6.1% share), 2025.
Security Boulevard, Hugging Face Has Become a Malware Magnet, 2024.10.23.
Search Engine Journal, Google Claims AI Overviews Monetize At Same Rate As Traditional Search, 2025.
SiliconFlow, OpenAI's gpt-oss Now Live: Designed for Agentic Workflows, 2025.
SQ Magazine, ChatGPT vs. Google Gemini Statistics 2025 (AI Overviews reach 2B users), 2025.
Stanford HAI, The 2025 AI Index Report, 2025.
Synced Review, Breaking LLMs' Limits: Upstage AI's SOLAR 10.7B Shines Bright with Simple Scaling Magic, 2023.
TechCrunch, Meta releases Llama 4, 2025.
TechCrunch, Google's Gemma AI models surpass 150M downloads, 2025.
TechCrunch, Sam Altman says ChatGPT has hit 800M weekly active users, 2025.
TechCrunch, Stability AI, Hugging Face and Canva back new AI research nonprofit, 2023.
TechTarget, Baidu makes foundation model Ernie 4.5 open source, 2025.
TechTarget, OpenAI o3 and o4 explained: Everything you need to know, 2025.
TechXplore, OpenAI chief says it needs new open-source strategy, 2025.
Tech Wave Arena, AI QA Automation: Reducing Costs and Time, 2024.
The Elec, NCSoft accelerates AI solution business with Llama-VARCO, 2024.
The Information, OpenAI COO Says ChatGPT Passed 11 Million Paying Subscribers, 2024.
The Reach, Huawei's New Chip Might Be The Companion DeepSeek Needed (PTX & Ascend), 2025.
Towards Data Science, DeepSeek-V3 Explained: Multi-Head Latent Attention, 2025.

Thunderbit, DeepSeek Statistics 2025: 33.7M MAU & $5.6M training cost, 2025.
Vizuara, Decoding Multi-Head Latent Attention (MLA): Reducing KV Cache by 93%, 2025.
Weights & Biases (W&B), Mistral AI Debuts Magistral: A Reasoning-Centric Language Model, 2025.
WSJ, Mistral AI Doubles Valuation to $14 Billion With ASML Investment, 2025.
XiaomiTime, Xiaomi announced official DeepSeek R1 supported devices (HyperOS integration), 2025.

[해외 사이트]

arXiv, https://arxiv.org/
ATOM Project, https://www.atomproject.ai/
EpochAI, https://epoch.ai/
EleutherAI, https://www.eleuther.ai/
GitHub, https://github.com/
Google Cloud, https://cloud.google.com/
HuggingFace, https://huggingface.co/
Linux Foundation, https://www.linuxfoundation.org/
Meta, https://www.meta.com/
OECD, https://oecd.ai/
OpenAI, https://openai.com/
OSI, https://opensource.org/
Star History, https://star-history.com
Wikipedia, https://en.wikipedia.org

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

ISBN 979-11-91525-73-1 (PDF)