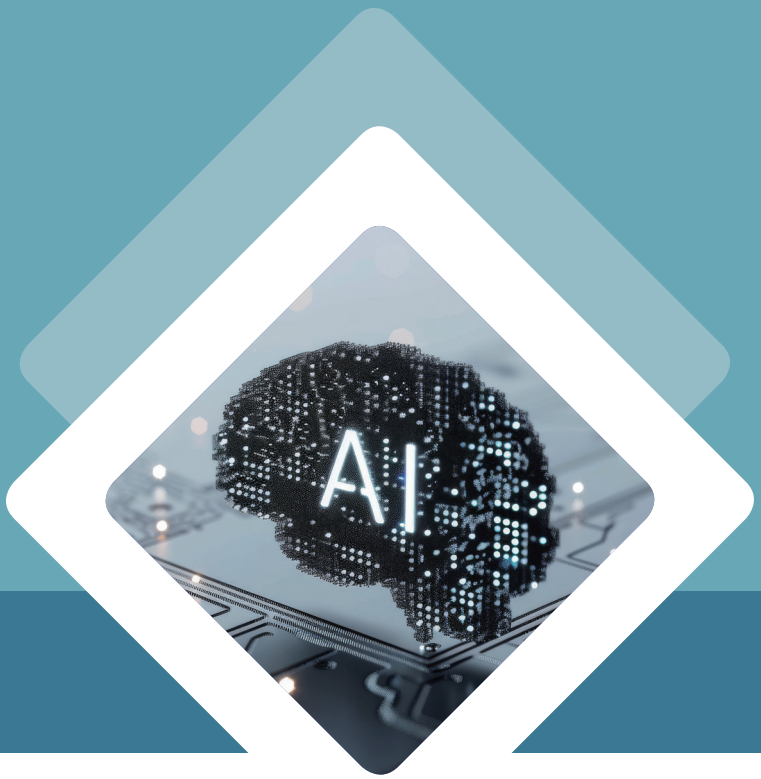
글로벌 초거대 AI 모델 현황 분석 (2020~2023년)

An Analysis on the Global Large-scale AI Model (2020-2023)



Executive Summary

인공지능(AI) 기술은 급격한 속도로 발전해왔으며, 특히 2020년대에 들어서면서 초거대 AI 모델이 경쟁적으로 등장하고 있다. 여기서 초거대 AI 모델은 대용량 연산 인프라를 바탕으로 방대한 데이터를 학습해 인간처럼 종합적인 인지·판단·추론이 가능해진 '큰 규모'의 AI 모델을 의미한다. 특정 목적에 따라 개별의 데이터를 수집·학습하여 만들어지는 기존의 일반 AI는 학습된 과업(Task)에 한하여 수행이 가능한 반면, 초거대 AI는 더욱 복잡하고 광범위한 분야에서 과업을 수행할 수 있다.

본고에서는 2020년부터 2023년까지 전 세계에 출시된 초거대 AI 모델 현황을 분석하고, 글로벌 기술 동향과 트렌드를 살펴보았다. 구체적으로, 미국 민간 연구단체인 'EPOCH AI'가 최근 업데이트('24년 7월)한 초거대 AI 모델 현황 DB를 통해 데이터를 수집하고, 2020년부터 2023년까지 출시된 초거대 AI 모델에 대해 출시년도, 국가, 분야, 과업유형, 개발형태, 개발조직 유형 등의 다양한 기준으로 정리·분석하였다. 우리나라 현황에 대해서도 주목하고, AI 분야에 대한 정책적 시사점을 도출하였다.

Artificial intelligence (AI) technology is developing at a rapid pace, and large-scale AI models are emerging as a competitive force, especially in the 2020s. These large-scale AI models have learnt vast amounts of data based on a large-capacity computational infrastructure, enabling comprehensive cognition, judgement, and reasoning akin to humans. Existing general AI, which is created by collecting and learning discrete data for a specific purpose, can only be applied to the tasks it has learnt, whilst large-scale AI can be applied to various tasks.

소프트웨어정책연구소 AI정책연구실 봉강호 선임연구원 bk91@spri.kr

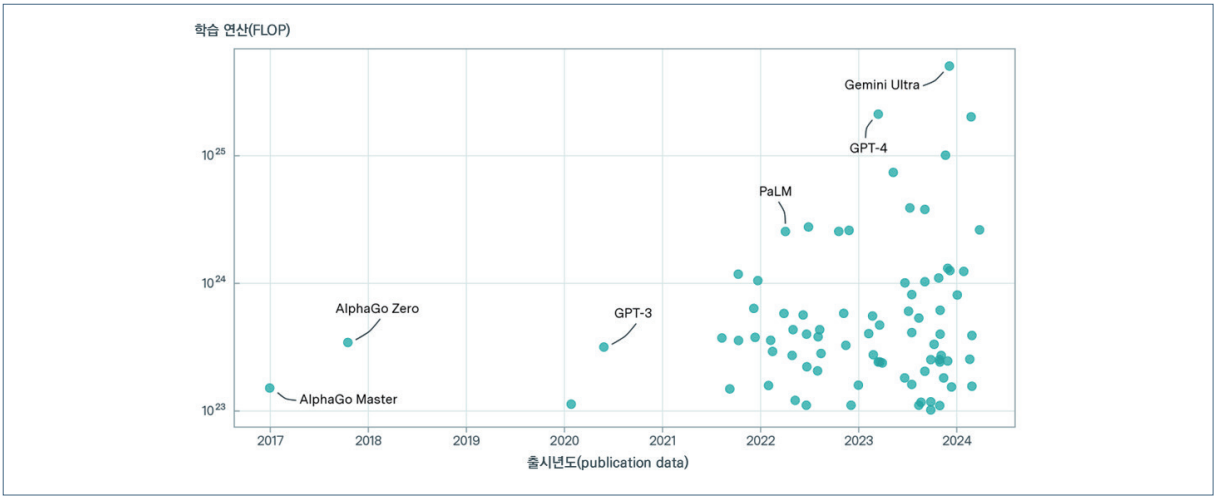
This report examines the global technology trends of large-scale AI models released between 2020 and 2023. Statistics on global large-scale AI models are available by release year, country, domain, task, development type, and organisation type. It gives special focus to the status of South Korea and provides policy recommendations for the field of AI.

Ⅰ. 서론

■ 연구 배경

- 인공지능(AI) 기술은 급격한 속도로 발전해왔으며, 특히 2020년대에 들어서면서 초거대(Large-scale) AI 모델이 경쟁적으로 등장함
  - AI 모델 성능이 '모델의 규모(Model size)', '학습 데이터(Dataset)' 및 '학습 연산(Training compute)'의 양에 따라 향상되는 경향성, 즉 '스케일링 법칙(Scaling law)'은 과거 AI 분야에서 가설에 불과했으나, 최근 이를 입증한 연구결과들이 발표됨(Henighan et al., 2020; Kaplan et al., 2020; Ghorbani et al., 2021; Behri et al., 2024)
  - 이후 스케일링 법칙을 근거로 한 AI 모델의 대형화 및 초거대 AI 모델 개발 시도가 급격하게 촉발됨[1]

[그림 1] 주요 초거대 AI 모델 출시 현황



주: 학습 연산량이 10^{23} FLOP 이상인 초거대 AI 모델 중 상위 규모의 81개에 대해서만 표기한 그림임
자료: Rahman et al.(2024)

1 물론 초거대 AI 모델 등장·발전의 기저에는 전 세계적인 디지털 데이터 규모의 증가, 연산 인프라(컴퓨팅 자원)의 발전, 알고리즘의 고도화 등이 있음

- 초거대 AI 모델은 대용량 연산 인프라를 바탕으로 방대한 데이터를 학습해 인간처럼 종합적인 인지·판단·추론이 가능해진 '큰 규모'의 AI 모델을 의미함
  - 이는 '초거대 언어 모델(Large Language Model; LLM)'을 포함하는 개념으로(안성원 외, 2023), 시각, 음성, 생체신호 등의 다양한 유형 모델도 이에 포함됨
  - 특정 목적에 따라 개별의 데이터를 수집·학습하여 만들어지는 기존의 일반 AI는 학습된 과업(Task)에 한하여 수행이 가능한 반면, 초거대 AI는 더욱 복잡하고 광범위한 분야에서 과업을 수행할 수 있음

■ 연구 내용

- 본고에서는 2020년부터 최근까지 전 세계에 출시된 초거대 AI 모델 현황을 분석하여 글로벌 기술 동향과 트렌드를 살펴봄
  - 2020년부터 2023년까지 출시된 초거대 AI 모델에 대해 출시년도(Release year), 국가(Country), 분야(Domain), 과업유형(Task), 개발형태, 개발조직 유형 등의 다양한 기준으로 정리하고 분석함
  - 추가적으로, 2024년 출시된 주요 모델을 소개함(파라미터, 학습 데이터 및 연산 규모 등)

- 특히 우리나라 현황에 대해서도 주목하고, AI 분야에 대한 정책적 시사점을 도출하고자 하였음

■ 연구 방법 및 자료

- 본 연구에서는 美 민간 연구단체인 'EPOCH AI'가 최근 업데이트('24년 7월)한 초거대 AI 모델 현황 DB를 통해 데이터를 수집하고 분석하였음
  - EPOCH AI는 역사적·현대적 관점의 AI 발전 궤적 분석을 위해 AI 모델 현황 관련 DB를 구축하였으며, 저작권 표시(CCL) 조건 하에 누구나 무료로 사용·배포·재생산할 수 있도록 공개함
  - 스탠포드大 HAI 연구소에서 매년 발간하는 「AI Index」의 경우에도 EPOCH AI의 데이터를 통해 AI 모델 현황을 분석해오고 있음

- EPOCH AI의 초거대 AI 모델 현황 DB에서는 GPT-3 수준인 1023 FLOP[2] 이상 규모의 모델을 초거대 AI 모델로 식별함
  - 학습 연산량에 관한 정보가 공개되지는 않았더라도 1023 FLOP 이상일 것으로 추정되는 일부 모델의 경우에도 동 데이터베이스에 포함되어 있음[3]

---

2 FLOP(floating point operations)은 AI 모델의 연산량을 파악하는 단위임

3 추정방법에 관한 자세한 내용은 EPOCH AI(2024)를 참조하기 바람

- EPOCH AI에서 검색방법의 한계로 인해 일부 모델*이 누락되었을 가능성의 존재를 밝히고 있다는 점은 유의해야 함
  - 벤치마크, 리포지토리, 인터넷 검색(뉴스, 웹사이트), 연구소 발표, 기타 문헌 검색 등의 방법으로 정보를 수집하였다고 밝힘

    * 비영어권 국가의 모델, 개발주체가 기밀 유지를 위해 발표하지 않은 모델 등

Ⅱ. 글로벌 초거대 AI 모델 현황

1. 2020~2023년 현황

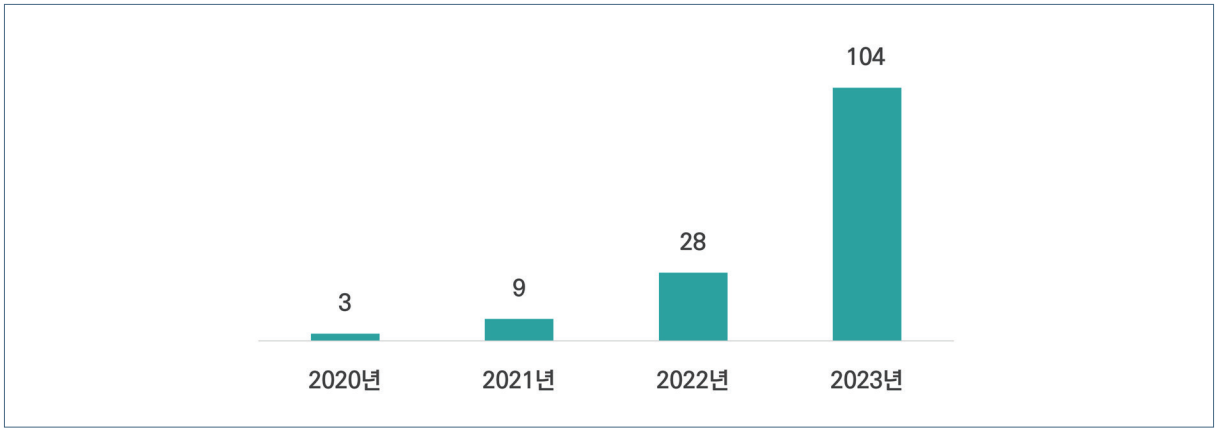
■ 종합

- 2020~2023년 기간에 전 세계적으로 총 144개(누적)의 초거대 AI 모델이 출시되었음

- 연도별로 보면, 2020년 3개의 모델이 출시된 이후 매년 급격하게 증가하는 추세임
  - 2020년 3개, 2021년 9개, 2022년 28개, 2023년 104개의 모델이 출시되었으며, 이는 연평균 226.1% 증가(CAGR)한 수준임

[그림 2] 연도별 초거대 AI 모델 출시 현황

(단위: 개)



자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

■ 국가별 현황[4]

- 2020~2023년 기간에 초거대 AI 모델을 가장 많이 개발한 국가는 미국(64건)으로 나타났으며, 그다음 중국(42건), 한국(11건), 프랑스(6건), 영국(5건) 순으로 많았음

- 2023년 단일년도 기준으로 보면 미국(41건), 중국(37건), 한국(8건), 프랑스(5건), 일본(3건) 순으로 많았음

[표 1] 국가별·연도별 초거대 AI 모델 개발 현황

(단위: 개)

국가명	모델 수					순위
	2020년	2021년	2022년	2023년	계	
미국	3	2	18	41 (중복 1)	64 (중복 1)	1
중국		2	3 (중복 1)	37 (중복 1)	42 (중복 2)	2
한국		3		8	11	3
프랑스			1	5	6	4
영국		1	3	1	5	5
일본				3	3	6
이스라엘		1		2 (중복 1)	3 (중복 1)	6
홍콩			1 (중복 1)	2 (중복 1)	3 (중복 2)	6
캐나다				2	2	9
독일			2		2	9
러시아			1	1	2	9
아랍 에미리트				2	2	9
핀란드				1	1	13
싱가폴				1	1	13

주: 중복은 타 국가와 공동개발한 모델이 중복 계산된 개수를 의미함
자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

- 2020~2023년 기간에 출시된 우리나라의 초거대 AI 모델은 총 11개로 조사됨
  - 네이버, 삼성, LG, KT, NC소프트, 코난테크놀로지 등의 기업이 개발한 모델임
  - 우리나라의 초거대 AI 모델은 언어(Language) 모델이 주를 이루고 있으며, 이 외에 이미지 생성(Image Generation), 비전(Vision), 바이오(Biology) 등의 모델이 있음

4 일부 복수 조직(국가)에 의해 개발된 모델을 중복 계산한 수치임

[표 2] 2020~2023년 출시된 우리나라 초거대 AI 모델 현황

출시년도	모델명	개발주체	분야	멀티모달 여부
2021	HyperCLOVA 82B	네이버	언어(Language)	
2021	HyperCLOVA 204B	네이버	언어(Language)	
2021	EXAONE 1.0	LG	언어(Language), 비전(Vision)	O
2023	EXAONE 2.0	LG	언어(Language), 이미지 생성(Image Generation), 바이오(Biology)	O
2023	VARCO LLM 2.0 base	NC소프트	언어(Language)	
2023	HyperCLOVA X	네이버	언어(Language)	
2023	Mi:dm 200B	KT	언어(Language)	
2023	Samsung Gauss Language	삼성	언어(Language)	
2023	Samsung Gauss Code	삼성	언어(Language)	
2023	Samsung Gauss Image	삼성	이미지 생성(Image Generation)	
2023	Konan LLM 41B	코난 테크놀로지	언어(Language), 비전(Vision)	

자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

■ 분야[5]별 현황

- 2020~2023년 출시된 초거대 AI 모델의 대부분은 언어 모델인 것으로 나타났으며, 그다음 비전, 이미지 생성, 비디오 순으로 많았음
  - 언어 모델(멀티모달 모델 포함)은 125개로 전체의 86.2%를 차지함
  - 2022년 이후부터는 비전, 이미지 생성, 비디오, 음성, 바이오, 로보틱스 등 다양한 분야의 초거대 AI 모델이 개발됨

[표 3] 분야별 초거대 AI 모델 현황(2020~2023년)

(단위: 개; 중복 포함)

분야	2020년	2021년	2022년	2023년	계
언어(Language)	3	9	23	90	125
비전(Vision)	0	1	2	12	15
이미지 생성(Image Generation)	0	0	3	10	13
비디오(Video)	0	0	2	5	7
음성(Audio & Speech)	0	0	1	4	5
바이오(Biology)	0	0	1	2	3
로보틱스(Robotics)	0	0	0	1	1

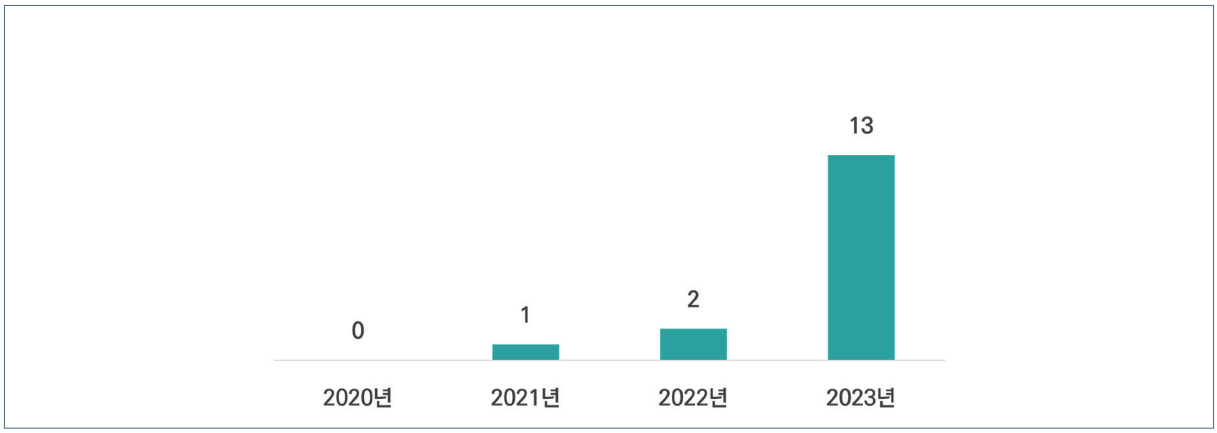
자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

5 EPOCH AI의 데이터베이스에서는 'Domain'이라는 용어를 사용하고 있는데, 이는 일종의 '모달리티(Modality; 즉, 데이터 형식/양식)'을 가리킨다고 볼 수 있음

- 멀티모달(Multimodal)[6]에 해당하는 초거대 AI 모델은 2023년까지 총 16개(11.0%)가 출시되었으며, 점차 많아지고 있는 추세임
  - 2021년 처음 등장하여 2022년과 2023년 각각 2개, 13개가 출시되었음
  - 여기서 2021년에 등장한 모델은 LG에서 개발한 EXAONE 1.0임

[그림 3] 연도별 초거대 AI 멀티모달 모델 출시 현황

(단위: 개)



자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

■ 과업(Task)유형별 현황

- 2020~2023년 출시된 144개 초거대 AI 모델 중 과업유형 관련 정보가 확인되는 모델은 총 129개였으며, 이 중 복수의 과업 수행 능력을 보유한 AI 모델은 51.9%(67개)의 비중을 차지하는 것으로 나타남

- 2020~2023년 출시된 초거대 AI 모델이 수행 가능한 과업유형 중 '언어 모델링/생성(Language Modeling/Generation)'이 가장 많은 것으로 나타남(87개 모델이 해당)
  - 그다음으로 채팅(Chat, 46개), 코드 생성(Code Generation, 25개), 번역(Translation, 19개), 질의응답(Question Answering) 순으로 많았음

6 멀티모달 AI 모델은 텍스트(언어) 데이터 외에도 음성, 이미지, 비디오, 생체신호(바이오) 등의 여러 데이터 유형을 처리·이해·생성할 수 있는 AI 모델을 의미함. 본고에서는 EPOCH AI의 데이터베이스상 멀티모달로 태그(지정)된 경우를 멀티모달 모델로 식별(계산)하였으며, 연구자가 임의로 멀티모달 모델 여부를 판단하지 않음

[표 4] 주요 과업유형별 초거대 AI 모델 현황(2020~2023년)

(단위: 개; 중복 포함)

과업유형(Task)	계	과업유형(Task)	계
언어 모델링/생성(Language Modeling/Generation)	87	비디오 생성(Video Generation)	5
채팅(Chat)	46	이미지 해석(Image Captioning)	4
코드 생성(Code Generation)	25	텍스트 요약(Text Summarization)	3
번역(Translation)	19	이미지 분류(Image Classification)	3
질의응답(Question Answering)	16	텍스트 기반 이미지 생성(Text-to-Image)	3
이미지 생성(Image Generation)	12	음성 인식(Audio Speech Recognition)	3
시각 질의응답(Visual Question Answering)	7	텍스트 자동완성(Text Autocompletion)	2

주: 일부 유사 과업유형들을 연구자가 통합하여, 2개 이상 모델이 해당되는 과업분야만을 표시함

자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

- 2022년을 기점으로 단일 모델로 수행 가능한 과업유형의 수가 크게 증가하는 추세를 보였으며, 이는 다중 과업에 적용가능한 모델 개발이 트렌드로 자리잡고 있음을 시사함
  - 두 가지 과업유형에 적용가능한 초거대 AI 모델은 2021년과 2022년 모두 한 자릿수만큼 출시되었으나, 2023년에는 32개로 크게 증가함
  - 세 가지 이상의 과업유형에 적용가능한 모델은 2022년 처음 출시(4개)되었으며, 2023년 더욱 많은(23개) 모델이 출시됨

[표 5] 과업유형 수별 및 연도별 초거대 AI 모델 현황(2020~2023년)

(단위: 개; 중복 포함)

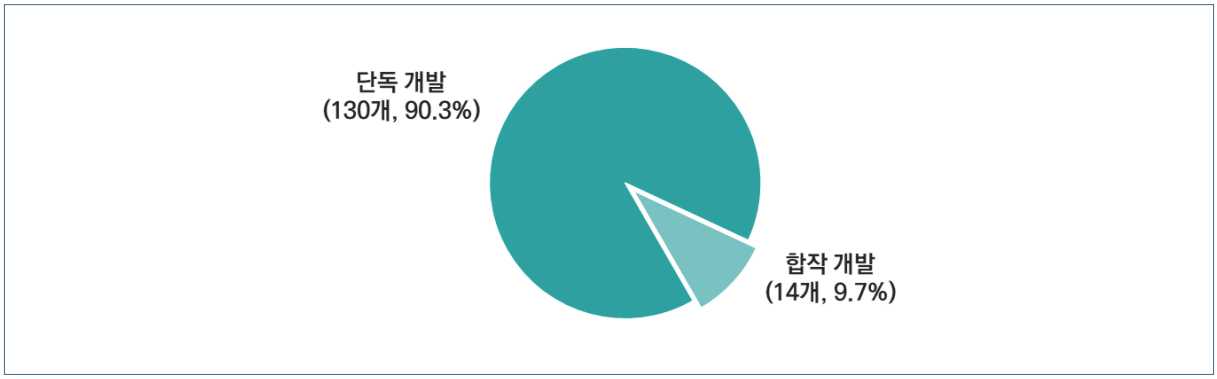
과업유형 수	2020년	2021년	2022년	2023년	계
1분야	2	3	16	41	62
2분야	1	3	4	32	40
3분야	0	0	4	8	12
4분야	0	0	0	12	12
5분야 이상	0	0	0	3	3

자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

■ 개발 형태 및 조직 유형별 현황

- 2020~2023년 출시된 초거대 AI 모델은 대부분 단일 조직이 개발(단독 개발)한 것으로 나타남
  - 전체 144개 모델 중 단독 개발 모델은 130개(90.3%), 복수 조직(기업)이 합작 개발한 모델은 14개(9.7%)로 확인됨

[그림 4] 개발 방식별 초거대 AI 모델 현황(2020년~2023년 누적)



자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

- 2020~2023년 출시된 초거대 AI 모델 개발조직은 대부분 산업계(기업)에 해당하는 것으로 나타남
  - 전체 144개 모델 중 134개(93.1%)가 산업계에서 단독 또는 합작 개발한 것으로 확인됨
  - 개발조직 유형이 정부/공공(Government)으로 확인되는 2개 모델은 아랍 에미리트(UAE)의 정부 출연연구기관인 Technology Innovation Institute (TII)에서 출시한 'Falcon'임
    * 2023년 출시(공개)된 Falcon-180B 및 Falcon-40B가 이에 해당됨

[표 6] 개발조직 유형별 초거대 AI 모델 현황(2020~2023년)

(단위: 개; 중복 포함)

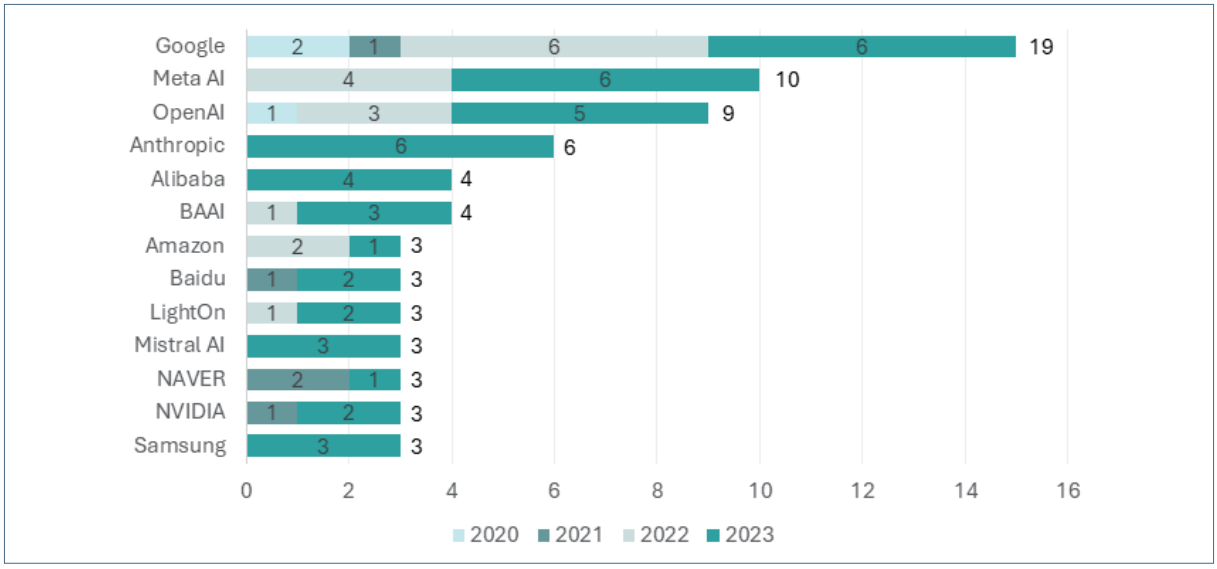
유형	2020년	2021년	2022년	2023년	계
산업계(Industry)	3	9	25	97	134
학계(Academia)		1	4	9	14
정부/공공(Government)				2	2
기타			1	1	2

자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

- 2020~2023년 기간 중 초거대 AI 모델을 가장 많이 출시한 조직은 Google(19개)인 것으로 나타남
  - 이는 DeepMind, Google DeepMind, Google Brain, Google Research 등 모회사인 알파벳 그룹의 조직을 모두 포함한 수치임
  - 이 외에 Meta, OpenAI, Anthropic, Alibaba, BAAI, Amazon, LightOn, Mistral AI, NVIDIA, 네이버, 삼성 등이 3개 이상의 초거대 AI 모델을 출시함

[그림 5] 3개 이상의 초거대 AI 모델 출시 조직

(단위: 개)



자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

2. 2024년 출시된 주요 모델

- 2024년에 출시된 초거대 AI 모델 중 학습 연산량(FLOP)을 기준으로 상위 4개의 모델을 선정하였으며, Llama 3.1-405B(Meta AI), Mistral Large (Mistral AI), Nemotron-4 340B(NVIDA), MegaScale(Production) (ByteDance 및 Peking University)이 이에 해당함
  - 분야(Domain)가 언어 모델에 해당하며, 학습 인프라로 NVIDIA의 제품을 사용했다는 것이 공통점임

[표 7] 2024년 출시된 주요 초거대 AI 모델

모델명	Llama 3.1-405B	Mistral Large	Nemotron-4 340B	MegaScale (Production)
분야(Domain)	언어	언어	언어	언어
과업(Task)	언어 모델링/생성	채팅	언어 모델링/생성, 채팅	언어 모델링/생성
매개변수(Parameters)	4.05E+11		3.4E+11	5.3E+11
학습 연산량(FLOP)	3.80E+25	2.00E+25	1.80E+25	1.20E+25
학습데이터 규모 (Datapoints)	1.56E+13		6.75E+12	
학습 인프라(HW)	NVIDIA H100 SXM5	NVIDIA H100 PCIe	NVIDIA H100 SXM5	NVIDIA A100
국가	미국	프랑스	미국	중국
개발조직	Meta AI	MISTRAL AI_	NVIDIA	ByteDance 北京大学 PEKING UNIVERSITY
출시일	2024-07-23	2024-02-26	2024-06-14	2024-02-23

주: 공개되지 않은 정보는 공란으로 표시함

자료: EPOCH AI(2024)의 데이터를 활용하여 연구자가 작성함

Ⅲ. 요약 및 정책적 시사점

■ 2020~2023년 글로벌 초거대 AI 모델 개발 트렌드 정리

- 2020~2023년 기간에 전 세계적으로 총 144개(누적)의 초거대 AI 모델이 출시되었으며, 2020년 3개의 모델이 출시된 이래 매년 급격하게 증가하는 추세임(연평균 226.1% 증가)

- 초거대 AI 모델의 대부분은 언어 모델이며, 이 외에 비전, 이미지 생성, 비디오, 음성, 바이오, 로보틱스 등 분야의 초거대 AI 모델은 2022년부터 본격적으로 등장함

- 2022년을 기점으로 △멀티모달을 지원하고 △다양한 과업을 수행할 수 있는 초거대 AI 모델이 빠르게 증가하여, 글로벌 개발 트렌드로서 자리잡고 있음
  - 초거대 AI 모델이 수행 가능한 과업분야로는 '언어 모델링/생성(Language Modeling/Generation)'이 가장 큰 비중을 차지하고, 이 외에 채팅(Chat), 코드 생성(Code Generation), 번역(Translation), 질의응답(Question Answering) 순으로 큰 비중을 차지하는 것으로 나타남

- 대부분 단독 개발 형태로 개발(단일 조직이 개발)되었으며, 산업계(기업)에서 주도해오고 있음
  * **(개발 형태)** 단독 개발 모델 비중 90.3% / **(조직 유형)** 산업계에서 개발한 모델 비중 93.1%

■ 우리나라의 초거대 AI 관련 글로벌 경쟁력 확인

- 우리나라는 미국과 중국 다음으로 세 번째로 많은 초거대 AI 모델을 출시하여, 글로벌 경쟁력을 일정 수준 이상으로 보유하고 있는 것으로 나타남

- 그러나 글로벌 초거대 AI 경쟁은 지속 심화되고 있는바, 우리나라가 경쟁력을 유지하고 선도국으로서 도약하기 위해서는 과감한 도전을 촉진하는 산업 생태계 육성 정책 추진 및 법·제도적 지원이 강조된다고 하겠음

■ 공동개발 장려를 위한 정책 필요

- 대다수의 초거대 AI 모델은 산업계(기업)로부터 개발된 것으로 나타났으며, 단독 개발(단일 조직이 개발) 형태가 주를 이루고 있음
  - 우리나라 초거대 AI 모델의 경우, 모두 산업계(기업) 및 단독 개발 형태에 해당함

- 우리나라가 글로벌 AI 개발 경쟁에서 뒤처지지 않기 위해서는 범국가적으로 자원과 역량을 결집하여 대응해나가야 함
  - 최근 초거대 AI 모델의 단점 및 한계점이 대두되는 동시에 온디바이스 AI[7]가 부상하면서 '소형 AI 모델'에 대한 관심이 급증하고 있으나, 규모가 큰 AI 모델의 중요성은 앞으로 더욱 높아지고 지속 강조될 것으로 전망됨
    * 대형 AI 모델의 단점 및 한계점으로는 △인프라·시스템 구축 및 운영에 수반되는 높은 비용(예: 데이터센터 구축, 전력 소모), △클라우드를 기반으로 AI 시스템을 운영함에 따른 민감 데이터/정보 보안성 관련 리스크, △외부 통신을 거치기 때문에 발생하는 대기시간(Latency)에 따른 느린 반응(결과물 산출)속도, △오프라인 환경에서의 이용 제약 등이 있음
    * 그러나 이러한 단점 및 한계점들은 앞으로 기술혁신을 통해 상당 부분 해소될 여지가 있는 만큼, △고지능·범용성을 보유한 초거대 AI 모델 시장과 △효율성·경제성 및 특정 영역에 대한 특화 가능성을 강점으로 가지는 소형 AI 모델 시장이 함께 성장하는 양상을 보일 것임
  - 이 중 초거대 AI 모델 개발에는 막대한 자금 및 자원(데이터, 인프라, 인력 등)이 소요되며, 현재 미국, 중국, 유럽 등 선진국에서는 천문학적인 규모의 투자가 이루어지고 있다고 알려짐(Maslej et al., 2024)
    * 스탠포드大 HAI 연구소에 따르면, 2023년 미국의 민간 AI 투자 규모는 약 672.2억 달러이며, 중국과 영국의 경우 각각 77.6억 달러, 37.8억 달러 규모임
    * 우리나라의 경우, 2023년 기준 민간 AI 투자 규모는 약 13.9억 달러로 전 세계 국가 중 9위 수준임
  - 이들 국가를 추격하는 입장에서, '머니 게임' 극복 방안이 필요한 상황임

- 따라서 기업 간 협력을 넘어서서, 생태계 차원에서 공동 대응할 수 있도록 유도할 수 있는 정책 마련이 필요하다고 판단됨
  - 중장기적인 개방형 혁신(Open innovation) 및 상생협력을 이끌어낼 수 있는 정책 수단을 개발하는 한편, 이를 저해할 우려가 있는 법·제도를 선제적으로 발굴·정비하는 것이 바람직함

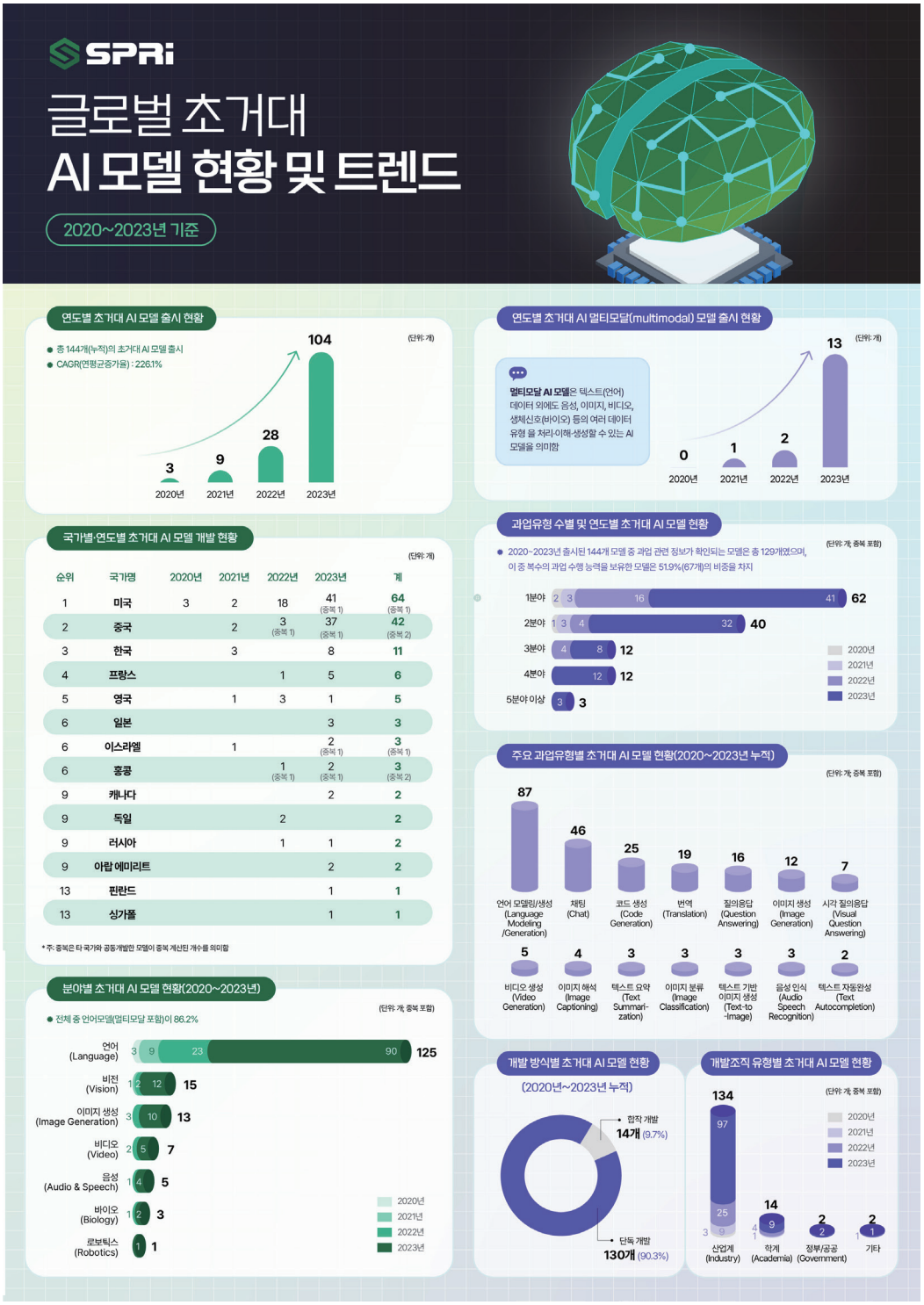
■ 글로벌 시장으로 진출할 수 있는 범용 모델 개발 지원

- 우리나라 초거대 AI 모델 중 다양한 언어를 지원하며 글로벌 시장에서 적용될 수 있는 공개된 범용 모델을 찾아보기 어려운 상황임
  - 물론 일부 예외(예: 네이버 HyperCLOVA X)는 존재하나, 대부분의 모델은 국내 서비스에 응용하거나 자사 제품 탑재를 위한 모델로 확인됨

---

7 온디바이스 AI는 노트북, 스마트폰, 태블릿 등의 기기(Device) 내에 탑재된 AI가 클라우드를 거치지 않고 자체적으로 작동하는 AI를 의미함. 클라우드 기반 AI에 대해 지속 거론되는 정보보안 측면에서의 한계점을 극복하고, 외부 통신 없이 기기 내부에서 AI 시스템을 구동하기 때문에 이용자가 보다 빠른 속도로 결과물을 얻을 수 있는 장점이 있다는 것이 특징임

- AI 분야에서 우리나라의 글로벌 영향력을 확대하기 위해서는 범용성 있고 공개 가능한 초거대 AI 모델 개발이 필수적이라 사료됨
  - 다양한 언어/분야에 적용가능한 범용 모델을 개발하는 국내 기업을 육성하거나 정부 주도로 개발을 추진하는 것도 하나의 방법이라 하겠음



⊙ 참고문헌

Behri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. (2024). “Explaining neural scaling laws”, Proceedings of the National Academy of Sciences, Vol.121, No.27, e2311878121, https://doi.org/10.1073/pnas.2311878121.

EPOCH AI (2024). “Large-Scale AI Models”, Published online at epochai.org. Retrieved from ‘https://epochai.org/data/large-scale-ai-models’ [online resource], (Accessed 24 July 2024).

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., Cherry, C. (2021). “Scaling Laws for Neural Machine Translation”, arXiv preprint, https://doi.org/10.48550/arXiv.2109.07740 (Accessed 24 July 2024).

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. (2020). “Scaling Laws for Autoregressive Generative Modeling”, arXiv preprint, https://doi.org/10.48550/arXiv.2010.14701 (Accessed 24 July 2024).

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). “Scaling laws for neural language models”, arXiv preprint, https://doi.org/10.48550/arXiv.2001.08361 (Accessed 24 July 2024).

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. (2024). *The AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April.

안성원, 유재흥, 조원영, 노재원, 손효현 (2023). “초거대언어모델의 부상과 주요이슈 – ChatGPT의 기술적 특징과 사회적·산업적 시사점”, 이슈리포트 IS-158, 성남: 소프트웨어정책연구소.