Journal of KIIT. Vol. 23, No. 1, pp. 191-199, Jan. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.1.191

AI 규제 정책을 통한 인공지능 기술도입 방향에 관한 연구 - 대규모 언어모델(LLM)을 중심으로

윤 철 희*

Research on the Direction of Artificial Intelligence Technology Introduction through AI Regulation Policy - Focusing on Large-scale Language Models(LLMs)

Cheolhee Yoon*

본 연구는 한국연구재단을 통해 과학기술정보통신부의 「해외우수과학자유치사업」의 지원을 받아 수행되었음(RS-2023-00304286)

요 약

본 연구는 대규모 언어 모델(LLM)의 기술적 발전과 각국의 AI 규제 정책을 분석하여 국내 도입을 위한 새로운 규제 프레임워크를 제시하였다. 최근 LLM 관련 주요 연구논문 분석을 수행하여 기술발전 동향을 체계화하였으며, 미국, EU, 일본 등 주요국의 AI 규제 정책을 비교 분석하여 정책적 시사점을 도출 후 이를 바탕으로 국내 환경에 적합한 원칙 기반의 LLM 도입 프레임워크를 설계하였다. 본 논문에서 제안하는 LLM 도입 프레임워크는 인간성 존중, 사생활 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성을 핵심 원칙으로 하며, 기존 AI 규제 프레임워크와 차별화되는 데이터 관리와 책임성 원칙을 통해 LLM 특성에 맞는 세부 규제방안을 제시하였으며, 특히 LLM의 건전한 발전과 안전한 활용을 위한 실천적인 규제 프레임워크를 제시함으로써, 국내 AI 생태계의 신뢰성 향상에 기여하였다.

Abstract

This study analyzed the technological development of Large-scale Language Models(LLMs) and AI regulatory policies of each country to propose a new regulatory framework for domestic introduction. In addition, we analyzed recent LLM-related research papers to systematize the trend of technological development. After drawing policy implications from a comparative analysis of AI regulatory policies in major countries such as the US, EU, and Japan, we designed a principle-based LLM introduction framework suitable for the domestic environment. The LLM introduction framework proposed in this paper is based on the core principles of respect for humanity, privacy, respect for diversity, non-infringement, publicness, solidarity, data management, and accountability, and presents detailed regulatory measures tailored to the characteristics of LLM through the principles of data management and accountability, which are differentiated from the existing AI regulatory framework, and is expected to contribute to improving the reliability of the domestic AI ecosystem by presenting a practical regulatory framework for the healthy development and safe utilization of LLM.

Keywords

large language motel, LLM, deep learning technologies, transformer models, artificial intelligence policies, ethical issues

* 경찰대학교 치안정책연구소 연구관
- ORCID: https://orcid.org/0000-0002-4862-4790

· Received: Oct. 11, 2024, Revised: Nov. 13, 2024, Accepted: Nov. 16, 2024

· Corresponding Author: Cheolhee Yoon
Police Autonomous Driving Center. Police Science Institute. Korea
Tel.: +82-42-968-2294, Email: bertter@police.ac.kr

Ⅰ. 서 론

인공지능 기술은 딥러닝 알고리즘과 컴퓨팅 파워의 비약적인 발전에 힘입어 최근 몇 년간 급속도로 진화하고 있다. 특히 자연어 처리(NLP, Natural Language Processing) 분야에서는 대규모 언어 모델(LLM, Large Language Models)이 등장하면서 새로운 패러다임을 맞이하고 있다. LLM은 GPT-3, BERT, XLNet 등 수십억 개의 매개변수를 가진 초대형 모델로, 방대한 텍스트 코퍼스를 사전학습(Pre-training)하여 인간과 유사한 수준의 언어 이해와 생성 능력을 보여준다.

LLM은 단순히 언어 처리 영역에 그치지 않고, 이미지, 영상, 음성 등 멀티모달 데이터와 결합하여 Caption 생성, 객체 인식, 음성 인식 등 다양한 태스크에 활용되고 있다. 또한 LLM을 기반으로 한 챗봇, 번역기, 컨텐츠 생성 도구 등 실제 서비스도 속속 등장하고 있다. 인공지능 모델의 대형화와 정교화로 성능과 활용성이 크게 향상되면서, 전 세계적으로 AI 기술에 대한 투자와 연구개발이 가속화되는 추세이다. 시장조사기관 보고서에 따르면 글로벌 AI 시장은 2025년까지 연평균 40% 이상의 고성장세를 이어갈 전망이며, 국내 AI 시장 역시 2030년까지 100조원 규모로 성장할 것으로 예측된다. 이처럼 AI 기술은 산업 전반의 효율성과 생산성을 높이는 핵심 동력으로 자리매김하고 있다.

그러나 LLM을 비롯한 AI 기술의 급속한 발전과 확산은 기술적, 사회적 우려를 낳고 있다. 알고리즘의 편향성과 차별, 윤리적 경계 모호성, 사이버 공격이나 가짜뉴스 생성 등 악용 가능성, 데이터 편중에 따른 지식 격차와 불평등, 개인정보 침해 및 프라이버시 문제 등 다양한 이슈가 제기되고 있다.

따라서 정부와 기업, 학계가 협력하여 AI의 개발과 활용 과정에서 발생할 수 있는 역기능을 최소화하고, 사회적 신뢰와 수용성을 확보하기 위한 법적, 제도적 장치 마련이 시급한 상황이다. 선제적인 규제와 정책을 통해 책임 있는 AI 생태계를 조성하고, 기술의 혜택이 사회 전반에 고루 돌아갈 수 있도록 하는 정책적 노력이 절실히 요구된다. 이에 본 논문에서는 LLM으로 대표되는 AI 기술의 동향과 사회경제적 파급효과를 분석하고, 기술 발전에 따른 사회적 쟁점과 해외 주요국의 정책 사례를 심층적으로 고찰하여, 이를 토대로 대규모 언어 모델(LLM) 기술의 건전한 발전과 안전한 활용을 뒷받침할 국내 정책 방향과 규제 프레임워크를 제안하는 것을 목표로 하였다.

Ⅱ. 대규모 언어모델의 기존연구

대규모 언어모델(LLM)의 발전과정을 분석한 최근 5년간의 선행연구를 살펴보면, T. B. Brown et al.[1]은 GPT-3를 통해 거대 언어모델의 Few-shot learning 가능성을 입증하였으며, 175B 파라미터 규모의 모델이 다양한 자연어 처리 작업에서 높은 성능을 보임을 검증하였다. J. Wei et al.[2]은 Instruction Fine-tuning이 언어모델의 Zero-shot 학습 능력을 크게 향상시킬 수 있음을 입증하였다. A. Chowdhery et al.[3]은 PaLM을 통해 540B 파라미터 모델의 추론 능력과 다국어 처리 성능을 검증하였고, L. Ouyang et al.[4]은 InstructGPT를 통해 인간 피드백 기반 강화학습(RLHF)이 모델의 유용성과 안전성을 개선할 수 있음을 보여주었다. 최근 Y. Liu et al.[5]은 대규모 언어모델의 지식 증류와 경량화 방법을 통해 연산 효율성을 높이면서도 성능을 유지할 수 있는 방안을 제시하였다.

언어 모델(Language model)은 주어진 텍스트 데이터를 기반으로 다음에 올 단어나 문장을 예측하는 모델로, 자연어 처리(NLP, Natural Language Processing) 분야의 핵심적인 요소 기술이다[6][7]. 그러나 전통적인 언어 모델은 범용성과 장기 의존성 문제로 인해 한계를 가지고 있었는데, 초기 NLP 연구는 주로 순환 신경망(RNN, Recurrent Neural Network) 기반의 Sequence-to-Sequence 모델을 활용하였다[8][9]. 그러나 RNN은 장기 의존성 문제, 즉 길이가 긴 문장에서 이전 단어의 정보가 손실되는 현상으로 인해 한계를 드러냈다. 이러한 문제를 극복하기 위해 2017년 Google에서 트랜스포머(Transformer) 아키텍처를 제안하였는데, 트랜스포머는 기존의 RNN 구조를 대체하여 Self-Attention 메커니즘을 통해 문장 내 단어 간의 상호 연관성을 학습한다.

이를 통해 정보 손실 없이 장문의 문맥 이해가 가능해졌다. 트랜스포머의 등장으로 NLP 분야에 새로운 패러다임이 형성되었으며, 이를 기반으로 다양한 사전학습 언어 모델(Pre-trained language model)들이 개발되었다[1][10] BERT(Bidirectional Encoder Representations from Transformers)는 트랜스포머 인코더를 활용한 양방향 언어 모델로, 문장의 양쪽 맥락을 모두 학습하여 언어 이해 능력을 크게 향상시켰다[11][12].

반면 GPT(Generative Pre-trained Transformer) 시리즈는 트랜스포머 디코더를 사용하여 문장 생성에 특화된 단방향 모델이다[13]. 이들 언어 모델은 기계독해, 문서 분류, 질의응답 등 다양한 NLP 태스크에서 최고 수준의 성능을 보여주며, 언어 모델의 범용성과 확장성을 입증하였다. 특히 GPT-3는 파라미터 수를 1750억 개까지 대폭 늘리고 Few-shot Learning 개념을 도입하여, 소량의 데이터로도 다양한 태스크 수행이 가능한 대규모 언어 모델(LLM)의 가능성을 보여주었다. 이후, GPT-3의 성공을 바탕으로 마이크로소프트의 Megatron-Turing NLG, 안트로픽의 Constitutional AI 등 더욱 규모가 크고 정교한 LLM들이 등장하였다[14][15]. 그리고 2022년 OpenAI가 공개한 ChatGPT는 GPT-3를 기반으로 강화학습을 접목하여 사용자와의 안전하고 효과적인 대화를 구현한 챗봇 서비스로, 사용 편의성과 활용성 측면에서 혁신을 불러일으켰으며 LLM 기술은 여전히 급속하게 발전 중으로 모델 경량화, 지식 증류, 수준 높은 추론 등 해결해야 할 과제와 함께 인공지능의 신뢰성 기술에 대한 고민을 하고 있다.

또한, 최근 몇 년간 대규모 언어 모델(LLM)이 자연어 처리 분야에서 괄목할 만한 성과는 이를 멀티모달로 확장하려는 시도로 이어졌다[16][17]. 2021년 OpenAI가 발표한 CLIP(Contrastive Language-Image Pre-training)과 DALL-E는 LLM과 이미지 데이터를 결합한 대표적인 사례로 CLIP은 이미지와 캡션 쌍을 대규모로 사전 학습함으로써, 이미지와 텍스트 간의 연관성을 학습하였다[18]. 이를 통해 주어진 텍스트에 맞는 이미지를 검색하거나, 이미지에 대한 설명을 생성하는 등의 태스크를 수행할 수 있게 되었다.

DALL-E는 CLIP에서 한 단계 더 나아가, 텍스트 입력을 기반으로 새로운 이미지를 생성해내는 모델이다. 이는 GPT-3와 같은 트랜스포머 언어 모델과, 이미지 생성에 특화된 VAE(Variational AutoEncoder) 등을 결합하여 구현되었다. 사용자가 원하는 콘셉트나 스타일을 텍스트로 입력하면, 해당 내용을 반영한 이미지를 자동으로 생성해주는 것이다. 구글의 Imagen, 바이트댄스의 ERNIE-ViLG 등 다양한 텍스트-이미지 생성모델들이 발표되며 관련 연구가 한층 더 가속화되는 추세이다.

최근에는 CLIP이나 DALL-E와 같은 멀티모달 모델을 활용하되, 텍스트 뿐 아니라 이미지, 오디오, 비디오 등 보다 다양한 미디어 형식을 아우르는 멀티모달 대형 언어 모델(MLLM, Multimodal Large Language Model)에 대한 연구도 진행 중이다[19]. 구글의 Gemini, OpenAI의 GPT-4, 마이크로소프트의 Kosmos-1 등이 MLLM의 대표적인 사례이다[20][21]. 이들은 방대한 멀티모달 데이터를 사전학습하고, 여러 태스크에 대한 few-shot 학습을 통해 이미지 캡셔닝, 시각적 질의응답, 텍스트-음성 변환 등 다양한 과업을 수행할 수 있다. MLLM은 입력 데이터의 형식에 제약받지 않고, 통합된 멀티모달 표현을 학습함으로써 인간에 가까운 수준의 '일반 지능'에 한 걸음 더 다가갈 수 있을 것으로 기대된다. 이는 검색, 챗봇, 콘텐츠 생성 등 다양한 AI 애플리케이션의 성능을 높이고 활용 범위를 넓히는 데 기여할 것이다. 또한 MLLM을 중심으로 한 AI 생태계를 선점하기 위한 빅테크 기업들 간의 경쟁도 가속화되고 있다.

III. 각국의 인공지능 규제 정책 고찰

미국은 인공지능 분야에서 글로벌 리더십을 유지하기 위해 정부 차원의 전략적 투자와 제도적 지원을 확대하고 있다[22][23]. 바이든 행정부는 인공지능을 국가 경쟁력의 핵심 요소로 인식하고, 연구개발(R&D) 투자 확대, 인재 양성, 윤리 및 거버넌스 체계 마련 등 다각도로 지원 정책을 추진 중이다[24][25]. 2021년 1월 '국가 인공지능 이니셔티브법(National AI Initiative Act)'이 제정됨에 따라 범정부

차원의 AI 전략을 총괄 조정하는 컨트롤타워로서 '국가 AI 이니셔티브실(NAIIO)'이 백악관 과학기술정책국(OSTP) 산하에 출범하였다. NAIIO는 연방 정부의 AI R&D 투자를 조정하고, 정부-산업-학계 간 협력을 촉진하며, AI 인력 양성과 인프라 구축을 지원하는 등 미국의 AI 경쟁력 확보를 위한 다양한 역할을 수행하고 있다.

또한, NAIIO는 AI 윤리 및 거버넌스 체계 수립을 주도하고[26][27] 미 국가과학기술위원회(NSTC)의 머신러닝-인공지능 소위원회를 통해 신뢰할 수 있고 책임감 있는 AI 개발을 위한 기본 원칙과 가이드라인을 마련하였으며, 연방 정부 차원의 AI 규제 프레임워크를 개발 중이다[28][29]. 특히 바이든 행정부는 알고리즘 편향성, 개인정보 보호, 설명 가능한 AI 등을 주요 이슈로 다루며 '인간 중심(human-centered)'의 AI 개발을 강조하고 있으며, AI 선도국으로서의 지위를 공고히 하기 위해 대규모 R&D 투자를 지속하고 있다. NSF는 AI 인재 육성을 위한 장학금 지원, 커리큘럼 개발, 교육 연구 등에 연간 6,500만 달러 이상을 투자하고 있으며, 노동부는 AI 및 자동화로 인한 일자리 대체에 선제적으로 대응하기 위해 직업훈련 및 전직 지원 프로그램을 대폭 확충하고 있다. 이처럼 미국은 강력한 AI 국가전략 추진 체계를 구축하고, 대규모 투자와 인재 양성, 글로벌 협력 등을 통해 AI 분야의 패권을 유지하기 위해 노력하고 있다.

표 1에서 언급하듯이 미국의 AI 연구개발 투자 및 실행계획 전략은 크게 세 가지 목표를 설정하고 있다. 첫째, 장기적 관점의 AI 연구를 지원하여 과학적 발견과 기술적 돌파구를 촉진한다. 설명 가능한 인공지능, 일반화 가능한 AI, 인간-AI 협업, AI 안전 및 보안 등이 주요 연구 주제로 거론된다. 둘째, 신뢰할 수 있고 윤리적이며 책임감 있는 AI 시스템을 개발한다. 이를 위해 AI 거버넌스 체계를 확립하고, 알고리즘 편향성 및 차별 해소, AI 윤리 교육 강화 등을 추진한다. 셋째, AI가 모든 분야에 활용될 수 있도록 AI 컴퓨팅 인프라와 데이터 접근성을 개선하고, 더 많은 AI 인재를 육성하는 것으로 특히 바이든 행정부가 강조하는 '인간 중심의 AI', '포용적이고 공정한 AI' 등의 가치를 반영하여 윤리 및 책임성 부문을 한층 강화한 것이 특징이다.

표 1. 미국의 AI에 대한 정부 거버넌스
Table 1. Government governance of AI in the US

Trait		User movie content evaluation points
Roles	Tissues	
General manager	Office of Science and Technology Policy (OSTP)	Overseeing and advising on federal AI activities; supporting interagency AI activities
Dedicate dagency	National AI Initiatives Office (NAIIO)	Implementing and coordinating the US AI strategy through cooperation with government departments and academia
supervisory	Special Committee on AI (SCAI)	Overseeing national AI initiatives and formulating AI R&D plans at the federal level
Advisors and councils	National Science and Technology Council (NSTC)	Advice on AI technology through affiliated technical committees
	National AI Advisory Committee (NAIAC)	Advice on AI R&D, ethics, standards, technology transfer, etc.
	AI National Security Council (NSCAI)	Advice on AI technology for national security and defense
	Work force Advisory Committee (AWPAB)	Advice on the adoption of AI in the labor market and the increase in automation
	National AI Research Resources TF (NAIRRTF)	Development and advice on the implementation roadmap for shared research infrastructure national AI research resources
study	National Science Foundation (NSF)	AI R&D and funding from major US universities and industries
guidelines and standard development	Machine learning and AI Subcommittee (MLAI-SC)	Develop AI standards and discover related requirements, and designate standard coordinators to support the execution of reliable AI strategies
Others/ defense	Joint AI Center (JAIC)	Department of Defense AI policy formulation, application of AI technology to military operations and tactics

기계학습, 자연어 처리, 컴퓨터 비전 등 AI 핵심 분야는 물론, 농업, 교육, 의료 등 AI 활용 분야까지 아우르는 종합적인 연구 체계를 갖추는 것을 목표로 하고 있으며, 미국은 중장기적 비전을 가지고 AI 연구개발에 전략적으로 접근하고 있다. 정부-산업-학계의 협력을 바탕으로 AI 기술과 인프라, 인력 등 AI 생태계 전반의 경쟁력을 높이는 한편, 사회적 신뢰와 수용성 확보를 위해 AI 윤리 및 거버넌스 체계 확립에도 공을 들이고 있으며, 미국의 경우 범국가적 차원의 AI 육성 전략을 토대로 글로벌 AI 패권 경쟁에서 우위를 점하고 있다. 유럽연합(EU)은 인공지능(AI)을 경제사회 전반의 혁신을 가속화하고 글로벌 경쟁력을 높이기 위한 핵심 기술로 인식하고, 2018년 4월 "EU를 위한 AI" 전략을 발표하며 본격적인 AI 정책을 추진하기 시작했다[30][31]. 동 전략은 AI 기술 및 산업 역량 강화, 사회경제적 변화 대비, AI 윤리 및 법적 프레임워크 마련이라는 세 가지 목표를 설정 후 2020년 2월에는 "AI 백서"를 통해 EU 차원의 AI 생태계 조성 방안을 제시하기도 하였다. 우수한 AI 생태계 구축을 위해서는 EU 및 회원국 간 협력, 산학연 협력, 중소기업 지원, 글로벌 협력 등을 통한 AI 연구개발 및 상용화 역량 강화가 필요함을 강조하며, 동시에 신뢰할 수 있는 AI 생태계 조성을 위해 인간 중심의 AI 개발 원칙을 수립, 알고리즘 투명성, 데이터 보호, 책임성 등 AI 윤리 이슈 해결을 위한 법적, 제도적 장치 마련이 시급함을 특히 강조하였다.

이러한 정책적 노력의 결실로 2021년 4월, EU 집행위는 세계 최초로 AI에 관한 법적 규제 체계인 "EU AI법(안)"을 발표하여, AI 시스템을 위험 수준에 따라 4단계로 분류하였다. 고위험 AI에 대해서는 시장 진입 전 적합성 평가, 사후 모니터링 등 엄격한 의무를 부과하였다. 특히 실시간 생체인식, 사회적 평가 등에 AI를 활용하는 것을 원칙적으로 금지하는 한편, AI 개발 과정에서의 편향성 검증 및 완화 조치, AI 시스템 등록제, 사고 발생 시 손해배상 등 전 주기적 규제 체계를 마련하였다. EU 주요국들도 자국의 강점을 살린 AI 국가전략을 수립하여 EU 정책에 부응하고 있는데, 대표적으로 독일은 2018년 "국가 AI 전략"을 발표하고, 2020년 및 2022년 업데이트를 통해 제조업, 의료, 모빌리티 등 주력산업 분야의 AI 활용 촉진에 주력하고 있다.

"AI Made in Europe" 기치 아래 EU 역내 국가들과의 협력을 통해 AI 기술 및 인프라 개발, 표준 수립, 스타트업 육성 등을 추진 중이고, 마찬가지로 AI 윤리규범 제정 등 AI 정책을 종합적으로 이행하기 위한 "AI 실행계획"을 마련 후 AI가 가져올 변화에 선제적으로 대비하고 사회 공론화 과정을 거쳐 정책을 수립함으로써, AI에 대한 시민사회의 신뢰를 확보하는 한편 AI 혁신을 뒷받침할 수 있는 제도적 기반을 다지고 있다. 일본 정부 경우 AI를 일본 경제와 사회 전반의 혁신을 가속화하고 국가적 과제를 해결하기 위한 핵심 수단으로 인식하고 있으며, 2019년 6월에는 범정부 차원의 "AI 전략 2019"를 발표하고, AI 기술 개발 및 사회 실장을 위한 중장기 로드맵을 제시하였다[14][15]. 동 전략은 일본의 미래상으로 제시된 "Society 5.0" 실현을 위해 AI 기술을 적극적으로 활용한다는 비전 아래, 교육, 의료, 교통, 방재, 인프라 등 다양한 분야에서의 AI 활용 방안을 모색하였다. 인간의 존엄성 존중, 다양성과 포용성 확보, 지속가능성 추구라는 세 가지 기본 이념을 바탕으로, 인재 육성, 산업 경쟁력 강화, 기술 체계 정비, 국제협력 확대라는 4대 전략 목표를 수립 후 "AI 전략 2022"이라는 AI의 윤리적·안전한 개발과 활용을 위한 정책 방안도 강조하고 있다. AI를 활용한 사회문제 해결과 경제 재도약을 위해 중장기적 관점에서 AI 정책을 수립·추진하고 AI 기술의 고도화와 더불어 실용화·산업화를 위한 정책적 지원을 병행함으로써, 일본 기업과 산업의 경쟁력을 한 단계 재고한다는 전략이다.

Ⅳ. 국내 인공지능 기술 적용 전략을 통한 대규모 언어모델(LLM) 적용 프레임워크

한국의 경우 2030년까지 한국을 세계 최고 수준의 AI 강국으로 도약시킨다는 목표 아래, 세계를 선도하는 AI 기술력 확보와 전 산업의 AI 융합 가속화 그리고 사람 중심의 AI 활용 확산이라는 3대 전략 방향을 설정하여 AI 기술을 활용한 주력산업 고도화, AI 활용에 대한 사회적 신뢰 구축, 미래 AI 인재 육성 체계 마련 등을 추진하고 있다. 앞선 선

도국들의 AI 기술 정책과 마찬가지로 한국은 AI 강국을 목표로 한 정부의 적극적인 육성책을 시도하고 있다. 하지만 핵심 원천 기술 부족, 전문 인력 부재, AI에 대한 사회적 수용성 문제가 있고 특히, AI 기술 발전이 가져올 윤리적·사회적 영향에 선제적으로 대응하기 위한 정책 보완이 필요한 시점이다[32]. 특히, 국내 인공지능과 대규모 언어모델(LLM)도입에 대해 개발 원칙을 수립하고, 알고리즘 투명성, 데이터 보호, 책임성 등에 대한 한국 중심의 AI 윤리 이슈 해결을 위한 법적, 제도적 장치 마련이 필요하다. 그림 1 LLM 프레임워크에서는 대규모 언어모델(LLM)을 기반으로 한 국내 도입 방향을 제시하고 있다[33].

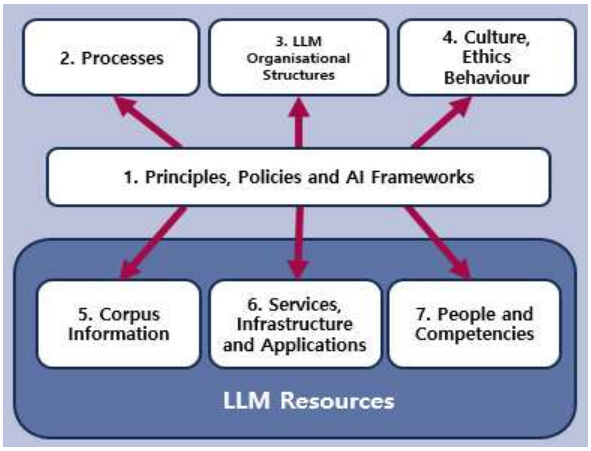

그림 1. LLM 프레임워크
Fig. 1. LLM Framework

대표적 인공지능기술로 분류되는 대규모 언어 모델(LLM)은 방대한 텍스트 데이터와 고성능 컴퓨팅 자원을 활용하여 인간 수준의 자연어 이해와 생성 능력을 구현한 AI 모델로, 현재 국내 산업계는 물론 일반 대중의 주목을 받고 있다. 한국 정부 역시 LLM을 차세대 AI 핵심 기술로 인식하고 집중 육성에 나서는 한편, 기술 발전에 따른 사회적 영향과 윤리적 리스크에 선제적으로 대응하기 위한 정책을 병행하고 있는 중이다. 결국 이러한 많은 관심과 활용으로 LLM의 사회적 영향력이 확대되고 있는 과정이며 이에 상응하는 윤리적 문제, 악용 가능성 등에 대한 우려도 연구되고 있다. 표 2 LLM 프레임워크 원칙과 같이 ①인간성 존중 ②사생활 보호 ③다양성 존중 ④침해금지 ⑤공공성 ⑥연대성 ⑦데이터 관리 ⑧책임성 등이 거버넌스와 병행으로 구성되어 LLM 개발 과정에서의 데이터 편향 방지, 개인정보 보호, 유해 콘텐츠 생성 억제 등을 수립해야한다. 그 기반으로 AI와 LLM 개발 시 전 단계에 걸친 LLM 프레임워크 체계를 마련함으로써 인간 중심의 신뢰할 수 있는 AI 생태계와 XAI의 사회적 수용 제고, 합리적 규제의 조화를 이루어 내어야 한다.

표 2. LLM 도입 프레임워크 원칙
Table 2. LLM framework principle

Principles	Illustrative
Respect humanity (development principles, policies)	LLM development guidelines, policies, and internal regulations
Transparent process (data protection)	Work execution procedures and performance roles
Culture, ethics and conduct (Respect diversity, prohibition of infringement)	Focus on individual and collective culture and ethics
Information reliability (Publicity)	Any information produced and used by the public
Services, infrastructure, and applications (data management, maintenance)	Management of applications and infrastructure used to provide IT-related services
People, skill expertise, and accountability	Competence to carry out all activities and decisions, and focus on accountability

Ⅴ. 결론 및 향후 과제

대규모 언어 모델(LLM)과 생성형 AI 기술의 급속한 발전은 자연어 처리를 비롯한 인공지능 분야에 새로운 패러다임을 가져왔다. 트랜스포머 아키텍처를 기반으로 한 LLM은 방대한 데이터로부터 인간과 유사한 수준의 언어 이해와 생성 능력을 학습함으로써, 기계번역, 문서 요약, 질의응답 등 다양한 애플리케이션에서 혁신을 이끌어내고 있다. 최근에는 LLM과 이미지, 음성 등 멀티모달 데이터를 결합한 MLLM까지 등장하며, AI 기술의 활용 범위와 파급력은 더욱 확대되는 양상이다.

그러나 대규모언어모델(LLM)으로 대표되는 AI 기술이 가져올 부정적 영향에 대한 우려인 알고리즘 편향성에 따른 차별 문제, 악의적 생성 콘텐츠 유포, 프라이버시 침해, 일자리 대체 등과 윤리적·사회적 리스크에 대한 대응 방안 마련을 고민하였다. 대규모언어모델(LLM)과 같은 대규모 블랙박스 모델의 투명성과 해석 가능성 확보가 기술 및 정책적 차원의 주요 화두가 되고 있으며, 이에 선도국들은 AI정책을 살펴보았다. 지속적으로 각국의 인공지능법 규제에 대한 연구실행과 인간중심적 연구의 수행을 통해 새로운 차원의 프레임워크 구성이 필요하며 이는 학문적, 실무적 연구가 통합된 AI체계로 발전해나가야 한다. 이를 위해 법학 뿐만 아니라 심리학, 인간공학, 인지공학, 윤리학 등 다학제간의 요소를 고려한 연구로 체계화 되어져야 한다. LLM으로 인해 가속화된 AI 기술의 발전은 산업과 사회 전반에 걸쳐 광범위한 혁신의 기회를 제공하고 있는 상황이고, 기술 발전이 개인의 권리 침해나 사회적 차별을 심화시키지 않도록 하는 안전장치 마련을 통해 기술의 혜택은 극대화하고 부작용은 최소화할 수 있도록 혁신 거버넌스를 꾸준히 모색해 나가야 할 것이다.

References

[1] T. B. Brown, et al., "Language Models are Few-Shot Learners", arXiv:2109.01652, pp. 1-75, May 2020. https://doi.org/10.48550/arXiv.2005.14165.

[2] J. Wei, et al., "Finetuned Language Models Are Zero-Shot Learners", arXiv:2109.01652, pp. 1-42, Sep. 2021. https://doi.org/10.48550/arXiv.2109.01652.

[3] A. Chowdhery, et al., "PaLM: Scaling Language Modeling with Pathways", arXiv:2204.02311, pp. 1-87, Apr. 2022. https://doi.org/10.48550/arXiv.2204.02311.

[4] L. Ouyang, et al., "Training Language Models to Follow Instructions with Human Feedback", arXiv:2203.02155, pp. 1-67, Mar. 2022. https://doi.org/10.48550/arXiv.2203.02155.

[5] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured Knowledge Distillation for Dense Prediction", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 45, No. 9, pp. 7035-7049, Jun. 2023. https://doi.org/10.1109/TPAMI.2020.3001940.

[6] DigitalDaily, "UK regulator unveils seven principles to stop a handful of big techs dominating AI", https://www.digitaldaily.co.uk/news/article/2023110401 [accessed: Nov. 04. 2023]

[7] Software Policy Institute, "The Rise of Generative AI and the Transformation of the Industry", SPRi Report, pp. 1-25, Mar. 2023.

[8] Software Policy Institute, "The State of the Generative AI Industry Ecosystem and Challenges", SPRi Report, Vol. 1, No. 2, pp. 1-30, Jun. 2023.

[9] Software Policy Institute, "Monthly AI Brief", SPRi Report, Vol. 2, No. 1-12, pp. 1-120, Dec. 2023.

[10] A. Vaswani, et al., "Attention is all you need", Advances in Neural Information Processing Systems, Vol. 30, pp. 5998-6008, Dec. 2017.

[11] R. Bommasani, et al., "On the opportunities and risks of foundation models", arXiv:2108.07258, pp. 1-192, Aug. 2021. https://doi.org/10.48550/arXiv.2108.07258.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?", ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623, Mar. 2021. https://doi.org/10.1145/3442188.3445922.

[13] J. Lee, "MUST HAVE Tencho's Pytorch Deep Learning Special Lecture", Gilbut, pp. 1-500, Sep, 2022.

[14] Korea Institute of Science and Technology Planning and Evaluation, "Analysis of The Global AI Index Results in 2023", KISTEP Report, pp. 1-45, Jun. 2023.

[15] Korea Advanced Institute of Science and

Technology, "Global AI Innovation Competition: Present and Future", KAIST Report, pp. 1-35, Dec. 2021.

[16] Korea Agency for Intelligence and Information Society, "2022 Information Society Statistics Collection", KAIS Report, Vpp. 1-80, Mar. 2023.

[17] Korea Institute of Public Administration, "Global Trends and Implications of AI Regulation", KIPA Report, pp. 1-50, May 2023.

[18] NIA, "Direction of Adoption of AI in the Public Sector Based on Large-scale Language Model", NIA Report, pp. 1-55, Apr. 2023.

[19] NIA, "Analysis of AI Strategies of Major Countries", NIA Report, Vol. 1, No. 2, pp. 1-60, Jun. 2023.

[20] D. A. Tamburri, "The AI Act: A Framework for Building Trustworthy AI Systems", Computer, Vol. 55, No. 10, pp. 15-23, Oct. 2022. https://doi.org/10.1109/MC.2022.3197119.

[21] D. Zhang, et al., "The AI index 2022 annual report", arXiv:2205.03468, pp. 1-222, May 2022. http://dx.doi.org/10.48550/arXiv.2205.03468.

[22] L. Floridi, et al., "AI4People—an ethical framework for a good AI society", Minds and Machines, Vol. 28, No. 4, pp. 689-707, Dec. 2018. https://doi.org/10.1007/s11023-018-9482-5.

[23] A. Jobin, M. Ienca, and E. Vayena, "Artificial intelligence: the global landscape of ethics guidelines", Nature Machine Intelligence, Vol. 1, No. 9, pp. 389-399, Sep. 2019. https://doi.org/10.1038/s42256-019-0088-2.

[24] J. Fjeld, N. Achten, and H. Hilligoss, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI", Berkman Klein Center Research Publication, Vol. 2020, No. 1, pp. 1-39, Jan. 2020.

[25] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines", Minds and Machines, Vol. 30, No. 1, pp. 99-120, Mar. 2020. https://doi.org/10.1007/s11023-020-09517-8.

[26] N. A. Smuha, "From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence", Law, Innovation and Technology, Vol. 13, No. 1, pp. 57-84, May 2021. https://doi.org/10.1080/17579961.2021.1898300.

[27] R. Chowdhury and S. Joshi, "Explainable artificial intelligence (XAI): An engineering perspective", Engineering Applications of Artificial Intelligence, Vol. 110, pp. 104629, Apr. 2022. https://doi.org/10.1016/j.engappai.2022.104629.

[28] A. B. Arrieta, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, Vol. 58, pp. 82-115, Jun. 2020. https://doi.org/10.1016/j.inffus.2019.12.012.

[29] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation", International Data Privacy Law, Vol. 7, No. 2, pp. 76-99, May 2017. https://doi.org/10.1093/idpl/ipx005.

[30] R. Geirhos, et al., "Shortcut learning in deep neural networks", Nature Machine Intelligence, Vol. 2, No. 11, pp. 665-673, Nov. 2020. https://doi.org/10.1038/s42256-020-00257-z.

[31] S. Larsson and F. Heintz, "Transparency in artificial intelligence", Internet Policy Review, Vol. 9, No. 2, pp. 1-16, Jun. 2020. https://doi.org/10.14763/2020.2.1469.

[32] K. Jang and W. Kim., ".Developing and Refining Components in Data Governance Framework", The Journal of Korean Institute of Information Technology, Vol. 14, No. 9, pp. 93-107, Sep. 2016. https://doi.org/10.14801/jkiit.2016.14.9.93.

[33] C. Moon and S. Kim, "A Study on Advanced Model for Personal Information Security Management", The Journal of Korean Institute of Information Technology, Vol. 13, No. 1, pp. 93-99, Jan 2015. https://doi.org/10.14801/jkiit.2015.13.1.93.

저자소개

윤 철 희 (Cheolhee Yoon)

2016년 8월 : 고려대학교 디지털포렌식학과(공학석사)
2023년 2월 : 연세대학교 기술정책(공학박사)
2024년 8월 : 극동대학교 인공지능보안학과(공학박사)
2017년 6월 ~ 현재 : 경찰대 치안정책연구소 연구관
관심분야 : 인공지능, 데이터분석, 딥러닝