Super-large AI and Generative Artificial Intelliger

Cho Young-im Head of the Korean Delegation, ISO/IEC JTC 1/SC 42, Professor in the Departme

1. Introduction

'super-large AI' refers to AI systems that have made significant advancements in scale, learning capabilities, and problem-solving abilities compared to typical artificial intelligence (AI) systems. This represents an evolved form of AI that can perform more complex and extensive tasks based on large amounts of data and processing power. Super-large AI primarily operates based on advanced technologies such as deep learning algorithms and reinforcement learning, enabling it to learn from large datasets, recognize patterns, and solve complex problems and make predictions. Super-large AI can be utilized in various tasks such as image recognition, natural language processing, and recommendation systems, achieving levels of performance similar to human learning, creativity, and problem-solving capabilities. These models are trained on massive datasets and require significant computational resources for learning and inference tasks. The GPT-4 model, announced on March 15, 2023, is a representative example. 'Generative AI' refers to AI that is specialized in generating texts, images, voices, etc. Here, 'generative' means a general-purpose AI that can autonomously create what the user requests without needing detailed instructions or training. Generative AI has the capability to generate and build new information, content, or data, and performs tasks such as prediction, creation, and modeling based on past data to produce new outcomes or solve problems. Generative AI can create new content from given inputs, generate images, or lead conversations, and is utilized in various areas such as natural language processing, image generation, voice generation, music composition, and artistic creation. Such models are commonly referred to as conditional generative models. The super-large AI model GPT-4 can leverage image data, and when images are inputted into GPT-4, it has the capability to create captions or classify and analyze images. In this way, super-large AI emphasizes scale and capability, while generative AI emphasizes the ability to create something new. Each has differing focal points, making it clearer to use them distinctly. Super-large AI exhibits enhanced learning and problem-solving abilities, with heightened predictive and performance characteristics. Generative AI, on the other hand, primarily utilizes generative modeling and natural language processing technologies to perform creative tasks. However, recently, there has been a tendency to generalize these as 'super-large generative AI', due to the close technological fusion and connection between these two. Thus, 'super-large generative AI' encompasses the characteristics and functions of both super-large AI and generative AI, covering large-scale data processing and generation tasks.

It refers to an integrated AI that performs AI systems.
In this paper, we will focus primarily on the field of super-large generative AI development and examine the major technologies and standard trends.

2. Major Technologies of Generative AI

2.1 Key Technologies

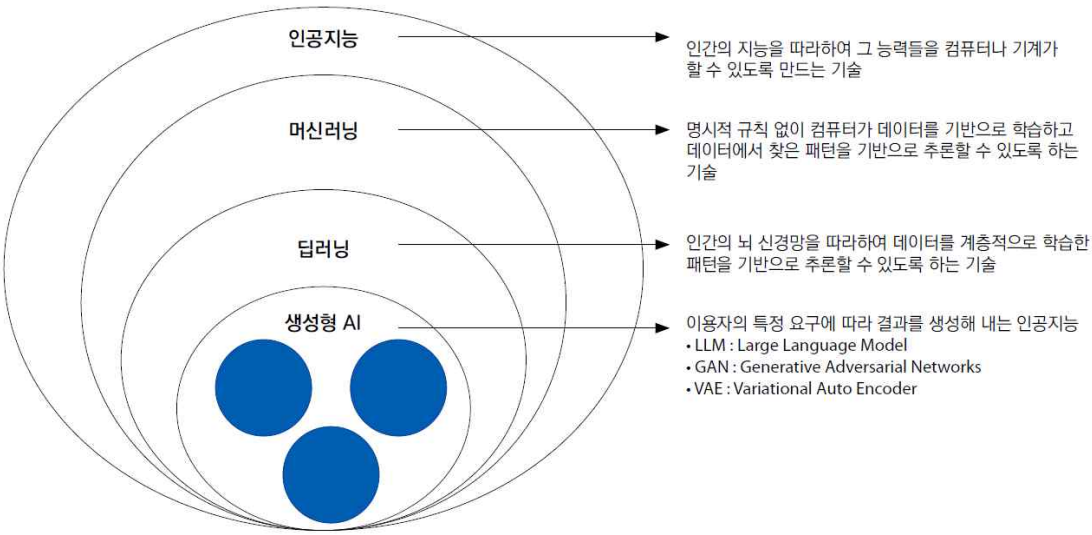
Generative AI refers to AI technologies that actively generate results based on specific user requirements. Until now, deep learning-based AI technologies were primarily focused on prediction or classification based on existing data, but generative AI represents a further evolution where AI autonomously seeks out and learns from data to actively provide results such as data or content in order to solve questions or tasks requested by users.

AI developers are creating and applying various generative AI models according to the purposes of the services they wish to develop. The most widely used generative AI model in chatbot services like ChatGPT is the Large Language Model (LLM). Simply put, the LLM is a generative AI model that learns language data such as text to provide results. The LLM applied to ChatGPT, developed by OpenAI, is GPT, and in March 2023, ChatGPT-4, which has about 500 times the model size of the existing model GPT 3.5, was released. Additionally, Google has unveiled its chatbot service 'Bard' utilizing its own LLM, PaLM (Pathways Language Model), while Meta has released the LLM called 'LLaMA (Large Language Model Meta AI)'. In South Korea, Naver has developed the 'OCEAN' super-large language model specialized for the Korean language and plans to launch the chatbot service 'HyperCLOVA X' based on OCEAN in July 2023 to compete with ChatGPT-4.

The relationships between AI, machine learning, deep learning, and generative AI are generally depicted in [Figure 1]. Initially, following the emergence of AI, feature extraction and classification were crucial for the machine learning phase, with these functions operating independently of each other. However, since the advent of deep learning, the construction of artificial neural networks, feature extraction, and classification have occurred not independently but as part of an integrated model, resulting in an intelligent system that produces outcomes based on hierarchically learned results. Recent advancements in the 2020s have led to the emergence of 'generative AI,' a subfield of AI technology that provides increasingly sophisticated services resembling genuine human creations.

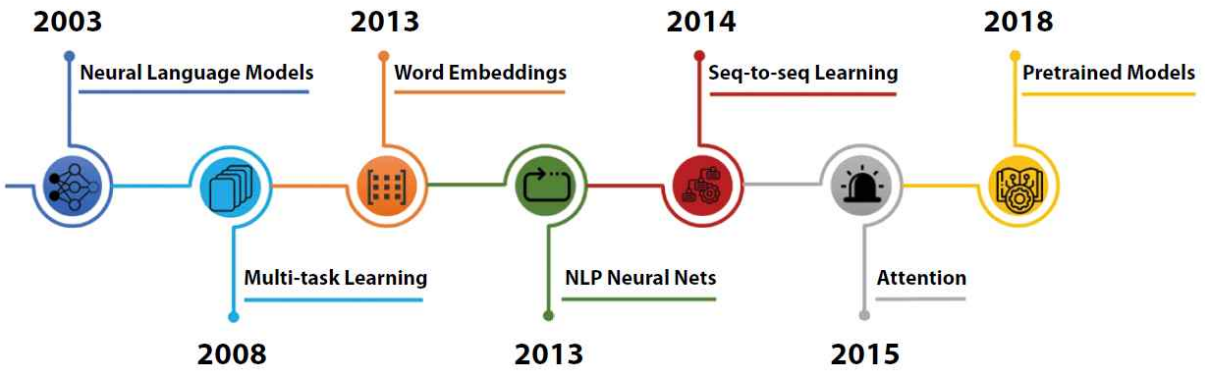
2.1.1 LLM (Large Language Model)

LLM stands for 'Large Language Model' and refers to large-scale language models. LLM refers to very large language models used in Natural Language Processing (NLP) and AI fields. These models can perform various NLP tasks such as language understanding, generation, and translation by learning from large amounts of text data. If they know the probabilities of words making up a sentence, they can effectively choose words or generate sentences sequentially by selecting the most plausible (highest probability) words among several candidates. This means they can understand and speak specific languages (e.g., English, Korean). Language models traditionally had very limited use due to the vast amounts of data they needed to collect and process to compute probabilities, but big data collection has made it possible to use them much more widely.



[그림 1] AI와 생성형 AI와의 관계

With the increase in computing power and infrastructure, LLMs have begun to receive attention as we enter the deep learning era. LLMs are currently drawing significant attention in the field of artificial int elligence, with prominent examples being OpenAI's GPT (Generative Pretrained Transformer) series a nd Google's BERT (bidirectional encoder representations from transformers). LLMs are based on mach ine learning and deep learning technologies, requiring large datasets for training and high computati onal capability. LLMs exhibit high performance across various natural language processing tasks and are applied in diverse areas such as text generation, machine translation, question-answering systems , chatbots, and summarization. LLMs can understand context based on trained data, providing approp riate responses or generation results, and can perform a variety of tasks related to language understa nding and generation. LLMs are considered a core technology for the advancement of NLP and for faci litating natural conversations between humans and machines. NLP is a technology that understands a nd generates human language, processing text data and inferring meaning. In generative AI, NLP tec hniques are utilized for tasks such as sentence generation, dialogue modeling, and translation. Transf ormer models and recurrent neural networks (RNNs) are primarily used for text generation. The follo wing [Figure 2] depicts the research trends from the early models of NLP to the present.

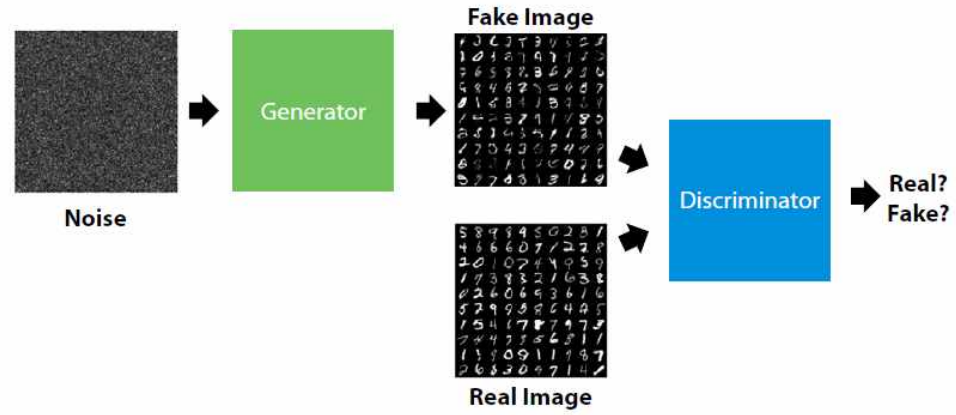


Source: A brief history of natural language processing-part 2, 2020

[Figure 2] Deep Learning NLP Technology Flow

LLM is classified into three types of technologies, similar to NLP. First, the language understanding mo del is a technology that pre-learns the context of words in natural language sentences to understand t he grammar and meaning of the words in the input sentence, with Google's BERT being an example. S econd, the language generation model is a technology that pre-learns natural language sentences an d predicts and generates the most suitable next word in the sequential order of the provided sequenc e of words, with examples including the GPT series, XLNet jointly developed by CMU and Google Brain, and Facebook's BART. Third, there are models that use both language understanding and generation together. This technology generates an output sentence based on the understanding of the input sent ence, outputting a sentence corresponding to the input sentence, with Google's T5 being an example.

2.1.2 GAN (Generative Adversarial Network) GAN is a deep learning-based model structure introduced by Ian Goodfellow in 2014 for generative modeling, consisting of two main components. One is the ge nerator, and the other is the discriminator. The generator tries to generate data similar to real data, w hile the discriminator tries to distinguish between the generated data and real data. These two eleme nts form a competitive and adversarial relationship that trains the model. Through this competitive an d adversarial learning, the generator generates data similar to the real data, and the discriminator im proves its ability to distinguish between fake and real data. [Figure 3] schematically represents the cor relation between the generator and the discriminator.

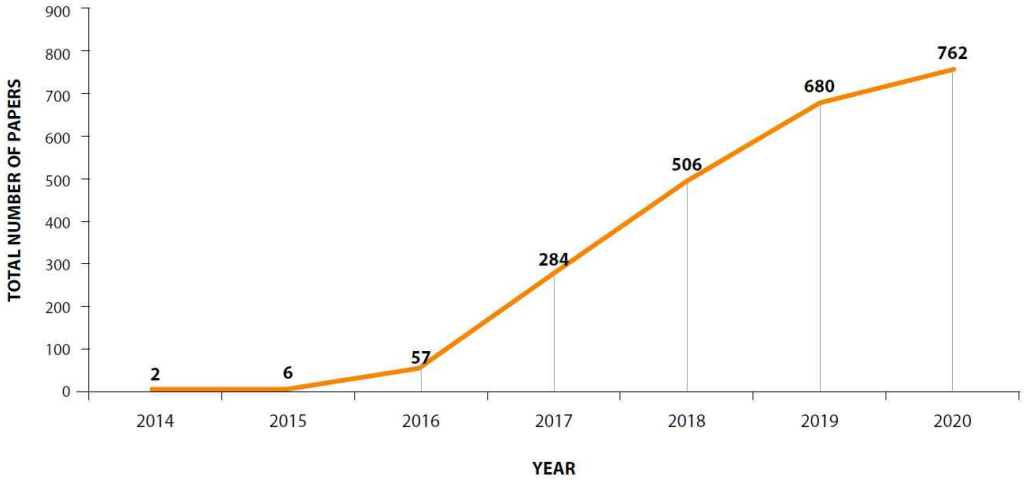


출처 : https://wikidocs.net/146217

[그림 3] 생성자와 판별자와의 상관관계

The following [Figure 4] shows the number of papers on GAN from 2014 to 2020, illustrating that consi derable research has been rapidly conducted. It is used in various application fields such as image gen eration, image transformation, and image sentiment analysis, and is a technology that greatly impacts AI research by generating new data.

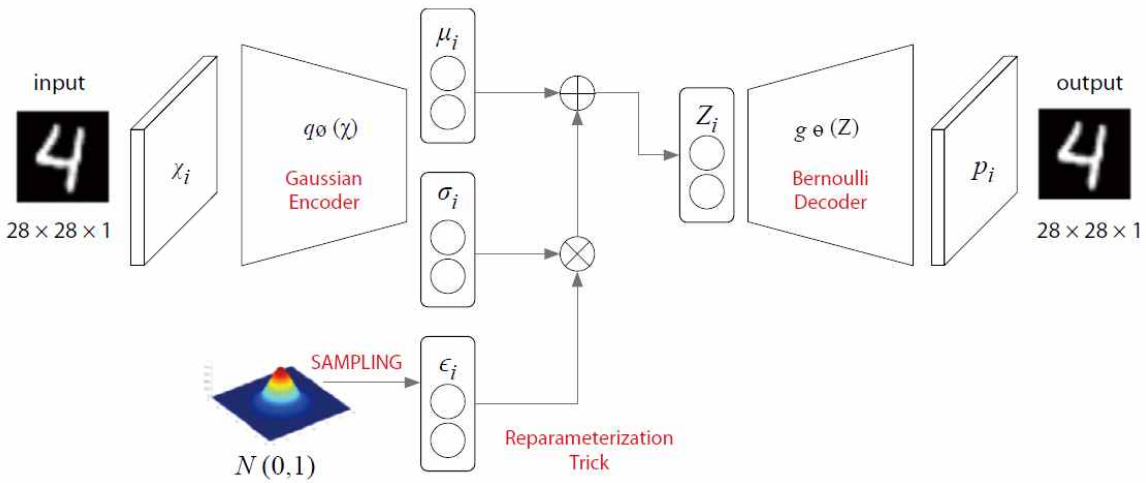
2.1.3 VAE (Variational AutoEncoder)
VAE stands for 'Variational Autoencoder', which means a variational autoencoder. VAE is one of the ge nerative models that learns the latent representation of the given data and uses it to generate new da ta.



출처 : https://wikidocs.net/146217

[그림 4] GAN을 이용한 논문 수

VAE is a model created by combining the ideas of deep learning and probabilistic modeling. VAE is based on an autoencoder structure that encodes the input data into a lower-dimensional latent space and adds probabilistic elements to the autoencoder structure that reconstructs the input data. VAE models a probability distribution to learn the latent representation of input data. The input data is represented as a probability distribution composed of mean and variance, and latent representations are generated through sampling from this probability distribution. The generated latent representations are used to represent the characteristics of the input data. The training process of VAE consists of minimizing the reconstruction error between the input data and the latent representations and a KL-divergence term that considers the distribution of latent representations. This allows the model to effectively learn the latent representations of the input data, thus enabling the generation of new data based on these representations. If the given training data p_data(x) has a certain distribution, it is basic to expect that the sample model p_model(x) will have the same distribution and that the inference value obtained from that model will be new data x[Figure 5].



Source: https://taeu.github.io/paper/deeplearning-paper-vae/

[Figure 5] Structure of VAE

VAE is used in various application areas such as image generation, image transformation, and latent representation learning. In particular, it is widely used along with GAN in image generation, possessing the property that moving in specific directions in the latent space causes images to transform. VAE can be effectively utilized in various generative modeling problems due to its ability to learn the latent representations of data and its probabilistic nature.

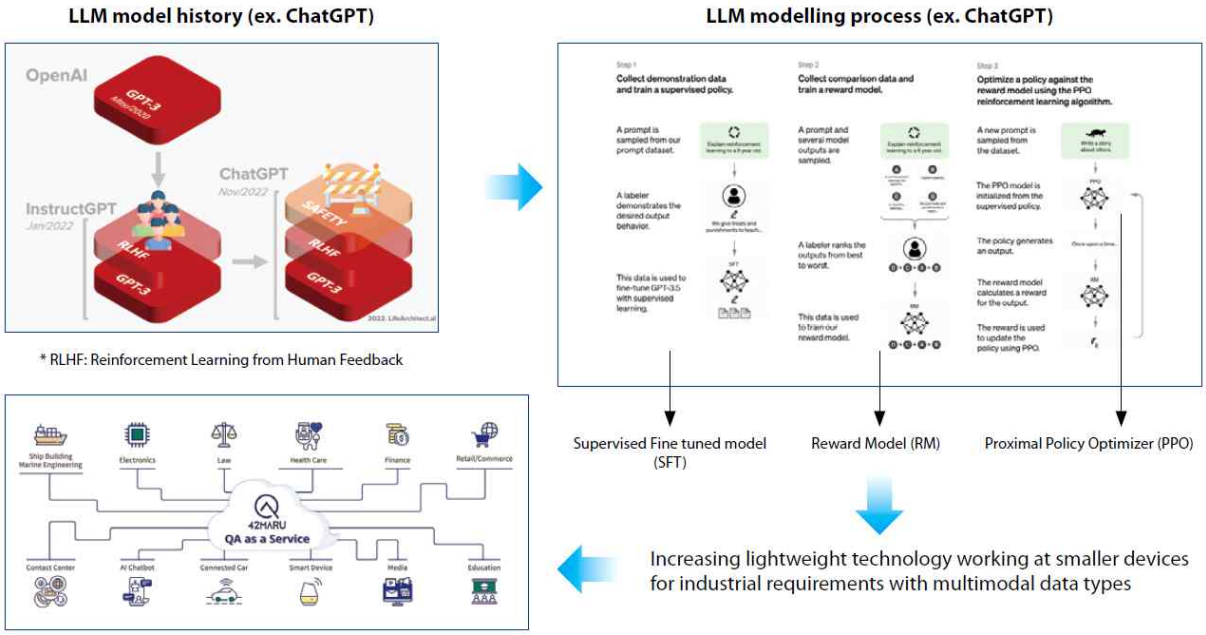
Let's take image generation, a representative application of VAE, as an example. Image generation is a technology by which generative AI creates and manipulates images. Models based primarily on Variational AutoEncoders and Generative Adversarial Networks (GANs) are commonly used in image generation. These models learn the characteristics of the given data and can generate new images based on this information. Music generation is a technology where generative AI composes and generates music. Music generation is based on algorithms that analyze given music data and understand the patterns and harmonies within the music. Recurrent neural network models such as LSTM (Long Short-Term Memory) are utilized for music generation, creating melodies, rhythms, and tunings. Video generation is a technology where generative AI creates videos, taking into account images and the passage of time to generate new footage. Generative modeling algorithms such as GANs and Variational AutoEncoders are used in video generation, ensuring continuity between video frames and creating natural movements.

Moreover, the technology of generative AI continues to evolve through advancements in data learning, model structures, and algorithms, contributing to the generation of creative works and the production of innovative content.

3. Trends in the Standard Development of Generative AI

There is currently no clear standard for generative AI. However, the need for standards is becoming evident. The most discussed standard is the LLM lightweight technology, which aims to reduce existing LLMs into smaller, lighter forms. This standard is being developed to reduce the size and computational costs of models, enabling efficient execution even in resource-constrained environments such as mobile devices or embedded systems. Recently, this has gained more attention in NLP. The concept of LLM lightweight technologies is illustrated in [Figure 6].

Standards are needed for LLM technologies due to resource limitations, deployment and dissemination, privacy issues, and collaborative research through interfaces. To overcome resource limitations, there is a need for the development of lightweight standards that allow LLMs to be scaled down and efficiently executed while considering resource constraints in mobile, embedded systems, or other limited resource environments. For efficient deployment and dissemination, lightweight standards need to be developed to enable shorter download, installation, and update times, even for smaller models. Additionally, since LLMs require large amounts of data, lightweight standards need to be developed to establish criteria for user privacy protection. For collaborative research through interfaces, it is essential to develop lightweight standards that provide an environment for sharing models and collaborating using standard interfaces and tools, thereby promoting technological advancement. To ensure that LLM lightweight models possess interoperability and productivity, the development of standardization technologies for frameworks and architectures is necessary. For example, frameworks, architectures, and interface standardization for LLM lightweight models should be developed.



출처 : OpenAI개요와 42MARU 인용하여 재구성

[그림 6] LLM 경량화 기술 개념도

The LLM lightweight standard will promote the dissemination and utilization of LLMs and can facilitate standardization and unity in the ecosystems of various companies and research institutions through resource efficiency, security, privacy protection, collaboration, and research support. Furthermore, the LLM lightweight standard could bring technical ripple effects such as interoperability between various applications, improved development productivity, enhanced performance, and strengthened security and reliability. Additionally, while not a direct standard for generative AI, the standards list being developed by the ISO/IEC JTC 1/SC 42 AI committee presents the standards being developed in WG 3 trustworthiness, as shown in <Table 1>. Here, numbers such as 20.00 indicate that when international standard development is approved and begins, it will start from 20.00, and when it reaches 60.60, it signifies that the international standard development has been completed and published.

<Table 1> Status of Development Standards from ISO/IEC JTC 1/SC 42 WG 3 Trustworthiness Working G

Document Number	Status	Document Title
ISO/IEC 23894:2023	Publication	Artificial Intelligence – Guidance on risk management
ISO/IEC TR 24027:2021	Publication	Artificial Intelligence – Bias in AI systems and AI aided decision making
ISO/IEC TR 24028:2020	Publication	Artificial Intelligence (AI) – Overview of trustworthiness in Artificial Intelligence
ISO/IEC TR 24029-1:2021	Publication	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
ISO/IEC FDIS 24029-2	50.20	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods
ISO/IEC PRF TR 24368:2022	Publication	Artificial Intelligence (AI) – Overview of ethical and societal concerns
ISO/IEC AWI TS 5471	20.00	Artificial intelligence – Quality evaluation guidelines for AI systems
ISO/IEC WD TS 8200	20.60	Artificial intelligence – Controllability of automated artificial intelligence systems

ISO/IEC 25059	60.00	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality Model for AI systems
ISO/IEC CD TR 5469	30.60	Artificial intelligence – Functional safety and AI systems
ISO/IEC CD TS 12791	30.60	Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks
ISO/IEC AWI 12792	20.00	Artificial intelligence – Transparency taxonomy of AI systems
ISO/IEC AWI TS 29119-11	20.00	Artificial intelligence – Testing for AI systems – Part 11:
ISO/IEC AWI TS 6254	20.00	Artificial intelligence – Objectives and methods for explainability of ML models and AI systems
ISO/IEC AWI TS 17847	20.00	Artificial intelligence – Transparency taxonomy of AI systems
ISO/IEC AWI TR 20226	20.00	Artificial intelligence – Environmental sustainability aspects of AI systems
ISO/IEC AWI TR 42106	20.00	Artificial intelligence – Overview of differentiated benchmarking of AI system quality characteristics
ISO/IEC AWI TR 21221	20.00	Artificial intelligence – Beneficial AI systems

4. Conclusion

Artificial intelligence is evolving rapidly and is becoming large-scale and convergent at an unprecedented pace. Therefore, the various plans that the government should set for technology development or workforce training have reached a level where the development details are too slow and outdated to keep up. Consequently, emerging technologies like artificial intelligence, which evolve daily, should be developed first by companies or institutions that recognize the issues, with the government promptly supporting the necessary parts. Of course, the government may also take the lead proactively, but this would be a challenging task as it relates to timing that requires reading the flow of the tech market. Generative AI is associated with super-large AI and several research topics. Super-large generative AI learns based on data collected on a large scale, but issues related to a lack of training data, excessive computing resource use, maintenance of relevance, and securing reliability are continuously being raised. Recently, research on data lightweighting technology and improvements in machine learning inference methods have been topics being investigated to solve these issues. The issue of unethical practices and bias removal technology in super-large generative AI models is also significant. As the use of super-large AI has become more widespread, issues of AI's unethicality and bias have been proliferating. This is due to the dependence on the training data, resulting in skewed content based on race, gender, and political inclinations. Recently, to address this, fairness assessment tools and frameworks have been widely researched to identify, classify, and remove the characteristics of data that can influence model bias. In particular, overseas, major IT companies are building systems to quickly address issues when they arise by publicly sharing models and data for non-commercial purposes, validating various aspects of the training data and models to mitigate unethical practices and biases. Research is also needed to prevent the generation of inappropriate responses caused by hallucination phenomena. As AI becomes commonplace, the issue of the reliability of results is emerging, and verifiable production must be ensured.

Interest in generative AI is growing, but solutions are not clear. Various technological attempts, such as masking techniques and proof tree construction, are being made to determine the factuality of results and to address the hallucination problem. In particular, methods like involving human feedback during the learning process or verifying generated results before providing the final output are being researched. Generative AI emphasizes the ability to create new things, while super-large AI emphasizes learning and problem-solving capabilities. In the future, artificial intelligence is expected to evolve into super-large generative AI, gaining the ability to both create new things and solve problems, thus gradually approaching what can be considered the ultimate goal of artificial intelligence: Artificial General Intelligence (AGI).

[References]

[1] Yang Ji-hoon, Yoon Sang-hyuk, Beyond ChatGPT in the Era of Generative AI: Media Content Generative AI Service Cases and Strategies for Securing Competitiveness, Media Issue & Trend, 2023 03+04 VOL. 55)
[2] Lim Soo-jong, Analysis of Trends in Super-large AI Language Models, ETRI, 2021
[3] Cho Young-im, Issues in Artificial Intelligence and Trends in International Standardization, Software Policy Research Institute, 2021 [4] A brief history of natural language processing-part 2, 2020
[5] https://taeu.github.io/paper/deeplearning-paper-vae/
[6] https://wikidocs.net/146217
[7] ICT Standardization Strategy Roadmap, TTA, 2023 (draft)
[8] Ministry of Science and ICT, Measures to Enhance Competitiveness of Super-large AI, 2023
[9] Ministry of Science and ICT, Promotion of Pilot Projects for Utilizing Super-large AI, 2022
[10] Ministry of Science and ICT, Strategies for Realizing Trustworthy Artificial Intelligence, 2021
[11] MT Report, The Era of Generative AI - Where is Korea Heading?, 2023

※ Source: TTA Journal No. 207