Editorial | Uncover This Tech Term

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.0790
Korean J Radiol 2023;24(10):1038-1041

Korean Journal of Radiology
KJR

Uncover This Tech Term: Foundation Model

Kyu-Hwan Jung[1,2]

[1]Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea
[2]Dataset Science Research Institute, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

Keywords: Foundation model; Artificial intelligence; Transformer; Large language model; ChatGPT; Representation; Few shot; Zero shot

What is the Foundation Model?

The foundation model (FM) is a family of machine artificial intelligence (AI) models that are generally trained by self-supervised learning using a large volume of unannotated dataset and can be adapted to various downstream tasks [1].

The most well-known examples of FMs are large language models (LLMs), such as ChatGPT [2]. Similar to ChatGPT, LLMs typically consist of billions of parameters and are designed to perform various natural language tasks. An LLM is initially pretrained to predict next words that follow a given input text (referred to as 'pretext task'), through which the LLM learns the semantics and structure of languages. With subsequent fine-tuning by human feedback, the LLM then acquires capabilities to generate natural and plausible responses to a wide range of queries (referred to as 'downstream tasks').

Training for the pretext task is typically achieved through self-supervised learning, using a massive unannotated dataset. The self-supervision dataset for the LLMs comprises a set of sentences with masked words generated by a computer without human involvement. The model is then trained to predict the masked words by looking at the unmasked part of the text. By utilizing vast amounts of self-supervised dataset, the model learns a semantically meaningful 'representation' (often colloquially referred to as feature) of the original dataset. These representations can be used for downstream tasks after fine-tuning with additional annotated datasets [3] or for generating new content in generative AI models [4].

'Transformers,' a special form of deep learning model, which can deal with a large volume of unstructured dataset [5] have been the architectural choice of many FMs. Transformers were originally developed for natural language processing to overcome the limitations of recurrent neural networks in solving sequence-to-sequence tasks. However, the expressivity and scalability of transformers based on self-attention mechanisms have rapidly expanded their application to other domains. Transformers applied specifically to computer vision are called vision transformers (ViTs) [6], and have shown remarkable success in improving the performance of convolutional neural networks.

Characteristics of FM

FMs demonstrate emergent abilities, which refer to the AI's capability to perform certain skills that were not explicitly intended during training, and these abilities improve as the scale of the model increases [7]. This is especially advantageous as it provides the potential for zero-shot or few-shot predictions in a novel task without additional training using an annotated dataset, which significantly reduces dataset construction efforts. Another characteristic of FMs is their ability to proficiently encode

Received: August 21, 2023 **Accepted:** August 23, 2023
Corresponding author: Kyu-Hwan Jung, PhD, Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, 115 Irwon-ro, Gangnam-gu, Seoul 06355, Republic of Korea
• E-mail: khwanjung@skku.edu

Korean Journal of Radiology
KJR

human instructions known as prompts and seamlessly incorporate them into their predictions [8]. This enables users to interact naturally with FMs through human-friendly prompts. FMs also have strengths in their capability to handle multimodal datasets. As transformers can flexibly handle various input types, we can efficiently build an AI model that can jointly learn the representation of inputs using a dataset from multiple sources [9]. The important concepts and terms related to FMs are listed in Table 1.

Types of FM

As explained earlier, LLMs are among the most actively developed FMs [2]. Another family of FM includes the vision or visual language foundation models (VLMs). VLMs learn the visual representations of objects from a large-scale image dataset and utilize these learned representations for downstream computer vision or vision-language tasks. The Segment Anything Model (SAM) [10] is a pioneering example of VLM developed to perform the pretext task of generating masks for all the objects in an image. The SAM can handle various types of prompts and demonstrates zero-shot generalization capabilities for novel objects or tasks. Contrastive language-image pretraining (CLIP) [11] is another VLM with the pretext task of aligning visual and textual representations for image and text pairs, and it demonstrates impressive zero-shot classification performance. There are other types of visual LLMs that utilize existing ViTs and LLMs as the encoder and decoder, respectively, and connect them using trainable alignment layers [12].

The most advanced and sophisticated type of FMs is the multimodal FM for integrating datasets from multiple domains or sources. These FMs enable AI to learn universal and robust representations of objects and concepts similar to the human cognitive system and are the most promising direction for implementing generalist AI [13].

Medical Applications

FM in medicine is currently one of the most active areas of AI research owing to its technically challenging nature and societal impact. While early studies focused on evaluating the performance of commercial general-purpose LLMs in clinical tasks [14,15], LLMs fine-tuned to the medical domain have been actively proposed [16-19] to further improve the performance in the clinical context. Variants of SAM [20] and visual LLMs [21,22] fine-tuned to the medical domain have been continuously introduced, and a CLIP-based VLM has also been proposed for zero-shot diagnosis of diseases in various modalities [23-25].

Recently, the first demonstration of multimodal medical FMs was proposed [26-28]. While these FMs are still in the proof-of-concept phase and mainly focus on interpreting medical images, they have shown promising capabilities for

Table 1. Terms related to foundation models

Term	Definition	Opposite/contrasting terms
Machine learning	A type of AI for developing systems that can learn or improve performance from based on the training dataset.	Rule-based system
Deep learning	A type of machine learning algorithms using artificial neural network with many layers to learn and utilize the representation of training dataset.	
Transformer	A type of deep learning model architecture with encoder and decoder blocks using self-attention mechanism to solve sequence-to-sequence problems.	
Generative AI models	A type of deep learning models trained to generate various types of contents conditionally based on the prompts.	Discriminative models
Large language models	A large-scale generative deep learning models trained to solve various natural language processing tasks.	Statistical language models
Generative pre-trained transformers (GPT)	A type of large language model using decoder part of transformer trained to predict or generate text based on the input.	
Self-supervised learning	A machine learning algorithm to learn the representation of dataset by generating self-supervision from the part of unannotated input dataset.	Supervised learning
Prompt or instruction	Additional queries given to generative AI models to generate desired output.	
Generalist or general AI (GAI)	An AI system that can learn from various type of dataset and solve wide range of novel tasks via adaptation.	Specialist or Narrow AI

AI = artificial intelligence

https://doi.org/10.3348/kjr.2023.0790

Table 2. Types of foundation models in medicine

Type	Reference FM	Medical domain or modality
BERT	NYUTron [3]	EHR
LLM	ClinicalGPT [16]	EHR, medical QA, medical dialogue
	RadiologyGPT [17]	Radiology report
	Med-PALM [18], Med-PALM 2 [19]	Medical QA
VLM	MedSAM [20]	General medical image
	SkinGPT-4 [21]	Skin image
	LLaVA-Med [22]	General medical image, Medical QA
CLIP	CheXzero [23]	Chest X-ray, report
	PLIP [24]	Digital pathology, description
	METS [25]	Electrocardiogram, report
Multimodal FM	BiomedGPT [26]	Radiology, pathology, dermatology, ophthalmology images, biomedical articles, EMR
	RadFM [27]	2D/3D radiology image-text pairs
	Med-PaLM M [28]	Radiology, pathology, dermatology images, genomics image, medical QA

FM = foundation model, BERT = Bidirectional Encoder Representations from Transformers, LLM = large language model, VLM = visual language foundation model, CLIP = contrastive language-image pretraining, EHR = electronic health records, QA = question answering, EMR = electronic medical records, 2D/3D = two-dimensional/three-dimensional

generalization to various high-performance clinical tasks. The various types of FM used in medicine and reference models are listed in Table 2.

Opportunities and Risks

FMs have immense potential to ease imminent healthcare problems that cannot be handled by conventional approaches using specialized or narrow AI models [1,29]. For patients, FMs would provide a natural and friendly interface and better access to clinical information, which reduces the knowledge gap. The versatile and interactive nature of FMs plays a critical role in assisting clinicians to make more confident and efficient decisions. For medical institutions, FMs would greatly reduce the burden of administrative or non-clinical tasks, which would facilitate the allocation of limited healthcare resources to unmet needs.

However, there are several challenges and risks that must be addressed before the deployment of FMs in clinical practice. In terms of technology, we need to better understand how FMs work, which will enable us to better utilize emergent abilities and prevent hallucinations generated by FM. From a clinical perspective, there are no established protocols or consensus on how to evaluate the efficacy and safety of FMs and how to regulate them, mainly because of the unlimited input and output of FMs [30]. For FM developers in the medical field, the integration of FMs with existing healthcare systems and the design of user interfaces that minimize misuse without compromising capabilities of FMs remain additional challenges.

Conflicts of Interest

The author has no potential conflicts of interest to disclose.

ORCID ID

Kyu-Hwan Jung
https://orcid.org/0000-0002-6626-6800

Funding Statement

None

REFERENCES

1. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv [Preprint]. 2021 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2108.07258
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv [Preprint]. 2020 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2005.14165
3. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. *Nature* 2023;619:357-362
4. Harshvardhan GM, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. *Comput Sci Rev* 2020;38:100285
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv [Preprint]. 2017 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.1706.03762
6. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words:

transformers for image recognition at scale. arXiv [Preprint]. 2020 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2010.11929
7. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv [Preprint]. 2022 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2206.07682
8. Gu J, Han Z, Chen S, Beirami A, He B, Zhang G, et al. A systematic survey of prompt engineering on vision-language foundation models. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2307.12980
9. Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. *Nat Commun* 2022;13:3094
10. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything [accessed on August 21, 2023]. Available at: https://segment-anything.com
11. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv [Preprint]. 2021 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2103.00020
12. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. arXiv [Preprint]. 2022 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2204.14198
13. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2303.03378
14. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology* 2023;307:e230987
15. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330:78-80
16. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.09968
17. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, et al. Radiology-GPT: a large language model for radiology. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.08666
18. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180
19. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2305.09617
20. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2304.12306
21. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. SkinGPT-4: an interactive dermatology diagnostic system with visual large language model. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2304.10691
22. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.00890
23. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng* 2022;6:1399-1406
24. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med* 2023 Aug 17. [Epub]. https://doi.org/10.1038/s41591-023-02504-3
25. Li J, Liu C, Cheng S, Arcucci R, Hong S. Frozen language model helps ECG zero-shot learning. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2303.12311
26. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2305.17100
27. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2308.02463
28. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2307.14334
29. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. arXiv [Preprint]. 2023 [cited August 21, 2023]. Available at: https://doi.org/10.48550/arXiv.2306.10070
30. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med* 2023 Jun 30. [Epub]. https://doi.org/10.1038/s41591-023-02412-6