

Foundation Models: From Current Developments, Challenges, and Risks to Future Opportunities

Ali Hussain*, Sikandar Ali*, Umm E Farwa*, Md Ariful Islam Mozumder*, Hee-Cheol Kim**

* Dept. of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

** College of AI Convergence, Institute of Digital Anti-Aging Healthcare, u-AHRC, Inje University, Gimhae 50834, Korea
 alihussainrana@gmail.com, sikandarshigri77@gmail.com, farwaalirana@gmail.com, arifulislamro@gmail.com
 heekii@inje.ac.kr

Abstract— Foundation models are a revolutionary technology in the field of artificial intelligence that could usher in an era of transformation in every field with its vision and language understating capabilities while maintaining remarkable performance. Here, we extensively review state-of-the-art foundation models in various fields such as LLMs, GPT, BERT, CLIP, etc. We also demonstrate some famous areas where foundation models are gaining much popularity, such as general foundation models, foundation models in the medical domain, education, law and finance, mathematics, autonomous driving, etc. These foundation models are trained on large-scale data by leveraging the capabilities of self-supervised learning approaches that ensure outperforming accuracy for relevant downstream tasks. They are successful due to pioneering architectural innovations, especially synergistically interleaving transformers and convolutional neural networks. These models have exhibited adaptability and resilience in various data patterns and conditions through sophisticated training paradigms such as self-supervised and supervised learning methods. While they hold transformative potential, they also have many challenges. The scale and quality of training data are still important predictors of model performance, and the need for interpretable and explainable AI systems becomes critically important. We highlight important research opportunities, high-performing computationally efficient and scalable architectures, and approaches to enable multimodal learning capacities. This review sheds light on the current state of foundation models and proposes a roadmap for their successful transitions.

Keywords— *Foundation models, self-supervised learning, vision-language models, LLMs, digital pathology foundation models,*

I. INTRODUCTION AND LITERATURE REVIEW

Nowadays, Foundation models (FMs) have been leading to breakthroughs in visual recognition, language understanding, and knowledge discovery. These models have shown tremendous performance in many areas. For example, in computational healthcare, FMs have been proven to possess outstanding logical reasoning and semantic capabilities that enable them to handle various forms of clinical data comprehensively [1-3]. This is evident in medical discussions, patient characterization, and program recommendations. Moreover, advanced handling capabilities provided by FMs for processing huge amounts of data are giving directions to analyzing real-life clinical data in a manner that was not

previously possible in healthcare routines, and the process is faster, too [4-8].

Indeed, one of the most prominent trends in FMs research is the focus on data-oriented strategies. First, FMs illustrate that the sizes of the model and data should be vast because the more information is received or processed, the greater the demand for significant training data [9, 10]. Second, FMs are homogenizing according to the results of their general applicability to downstream tasks. Consequently, the quality of data fed to the FM training plays an important role, both in the case of pre-trained FMs and the subsequent downstream models [11-13]. Therefore, resolving the main issues related to data is gradually unfolding as a high-priority focus area in research efforts [14, 15]. In the healthcare system, high-quality records could reveal all patient characteristics, including imaging, genomics, and lab testing data. Science-based solutions are supposed to revolutionize the client processes, assist in diagnostics, and provide fresh perspectives in therapy. However, some issues were and still are related to medical data, the data integration from multiple modalities, the dataset size, burdened data annotation, and the problem of patient data protection. There are new opportunities to develop data-oriented AI analytics in the foundation models (FMs) era. For example, multimodal FMs can provide large-scale approaches to integrating different data types. In particular, it is important to emphasize the great potential of FMs since they are capable of offering rather high-quality data, which may greatly contribute to the overall elimination of problems related to data deficit, abundance, and privacy in the sphere of medicine and healthcare [3, 16, 17].

II. FOUNDATION MODELS

A. General Foundation Models

1) Vision and Language Models:

Recently, there has been a lot of talk about foundation models. These are learning algorithms that are trained on large and generalized data to perform a variety of tasks. Unlike task-specific models, a foundation model can be considered the prior or the foundation of one or several other models. This is made possible by using transfer learning methodologies to either fine-tune or adapt a foundation model for a vast number of downstream tasks, using very little other data for each task. The most known types are the pre-trained large language models

(LLMs), such as Generative Pre-trained Transformers (GPT) [18]. OpenAI develops these language models. They are capable of writing text that would be indistinguishable from the text written by a human. They are trained on very large amounts of text and code. GPT-3 is a very knowledgeable language model that can write extremely well in many forms of text. GPT-4 is an advanced version that is better than GPT-3. It can process text and images. It is more reliable, creative, and better at executing complex instructions. BERT [19] is a language model created by Google. What's unique about it is the way it reads text. Other models read text one word at a time in sequence. BERT reads all of the words simultaneously. That allows it to interpret words based on all of the context around them, not just what appears before them. That method works incredibly well. T5 [20] is a text-to-text language model from Google AI. It converts all language tasks into a text-to-text paradigm. This includes treating such things as sorting text, answering questions, and translating similarly. This simplifies T5 to train and be proficient at a wide variety of tasks. RoBERTa [21] is by Facebook AI. It's really just BERT but improved. They trained it on more data and altered the way it learns. They also got rid of some components that didn't do much. These improvements cause RoBERTa to perform better than the regular BERT at most language tasks. Google and Carnegie Mellon University developed XLNet [22]. XLNet is a BERT variant. Instead of making predictions after masking words, XLNet makes predictions by taking into account all the possible permutations of the other words. This enables it to learn bidirectional context without some of the problems that BERT encounters. XLNet generally achieves state-of-the-art results on language benchmarks. DeepSeekR1 [23] is a language model from DeepSeek AI. It's good at coding and understanding language. It works well but doesn't cost too much to run. It comes in different sizes for different needs. DeepSeek AI wants to make AI open and available to more people. QwenVL [24] is made by Alibaba Cloud. It's a strong model that's open for anyone to use. It comes in different sizes and works with many languages. Alibaba Cloud made QwenVL open-source so more people can use and improve it. KIMI AI [25] is from a Chinese company called Kimi. Their models can handle really long text better than most others. This makes them good for summarizing documents, writing long content, and analyzing big chunks of text. KIMI AI is known for staying on topic even with very long text. Mistral AI [26] is a French company. Their models are smaller and faster than other top models but still work really well. They make their models open for anyone to use. Mistral AI models can handle many different language tasks. CLIP [27] is an OpenAI model that connects images and text. CLIP was trained using hundreds of thousands of pairs of images and text. It is capable of matching words with images very accurately. CLIP can detect objects within images that it never saw during training. It puts images and text into the same space so that similar ones gravitate closer to each other. Vision Transformer (ViT) [28] uses the Transformer model (for language) on images. Rather than considering pixels on a grid, it considers images in patches as one considers words in a sentence. It feeds patches into a Transformer. ViT demonstrated

that Transformers work well on images as well, not just text. SWIN Transformer [29] improves upon ViT and speeds it up. It performs better on more tasks such as object detection and image region sorting. SWIN employs a "shifted window" trick - it calculates small portions of the image independently but linked between layers. This performs better on finer image tasks. DALL-E [30] is another model from OpenAI. It generates images from text prompts. You type what you'd like to see, and it draws a picture. DALL-E trains a Transformer to understand the connection between text and pictures. It can generate creative, unexpected images from what you describe.

2) Vision and Language Models:

Wav2Vec 2.0 [31] is a speech model from Meta AI. It is trained on raw audio without up-front labels. This enables it to function with far less human-annotated data than previously. It trains first on unlabeled speech and then fine-tunes on a small labeled dataset. This is wonderful for languages that minimal training data exists to utilize. HuBERT [32] is a speech model from Meta AI. It operates by predicting covered segments of speech, i.e., filling in blanks. Rather than directly covering the sound waves, it covers hidden unit patterns that some other model first learns. This cleverness allows HuBERT to comprehend speech better. It performs effectively for speech recognition and other voice applications.

B. General Foundation Models

1) Vision and Language Models for Medical:

Several foundation models have been developed that have the potential to reshape the future of medical image analytics and diagnostics, as the field of computational pathology is advancing very quickly. UNI [18] and CONCH [19] are some such important developments in this regard. UNI has been trained on a colossal dataset of 115 million tissue patches and 100,000 whole-slide images and has demonstrated high performance on 34 clinical tasks ranging from disease detection to organ transplant analysis. CONCH takes a step further with 1.17 million pairs of histopathology images and text so that pathologists can recognize particular morphological features more accurately. EXAONEPath [20] is another significant model proposed for the processing of whole-slide images (WSI). The model contains 86 million parameters and was trained on 285,153,903 patches from 34,795 WSIs. State-of-the-art stain normalization methods are applied in the model to mitigate color variability and enhance performance across variable imaging conditions. It has been shown to surpass other models on linear evaluation tasks, which reflects its usefulness in pathology. Lunit [21] is an AI-powered tumor microenvironment analysis model. It was trained with the help of more than 150 pathologists, 50,000 slides, and more than 10 million single-cell annotations. It offers cancer, stroma, and tumor-infiltrating lymphocyte insights for 16 cancers. As a multimodal generative foundation model for slide-level histopathology, which we call the PRISM [22] Pathology model (Pathology Representation with Image-based Slide-level Modeling). An image found out of challenge is a multimodal generative foundation model designed to examine entire slide images at the slide stage, reflecting how medical pathologists method the situation.

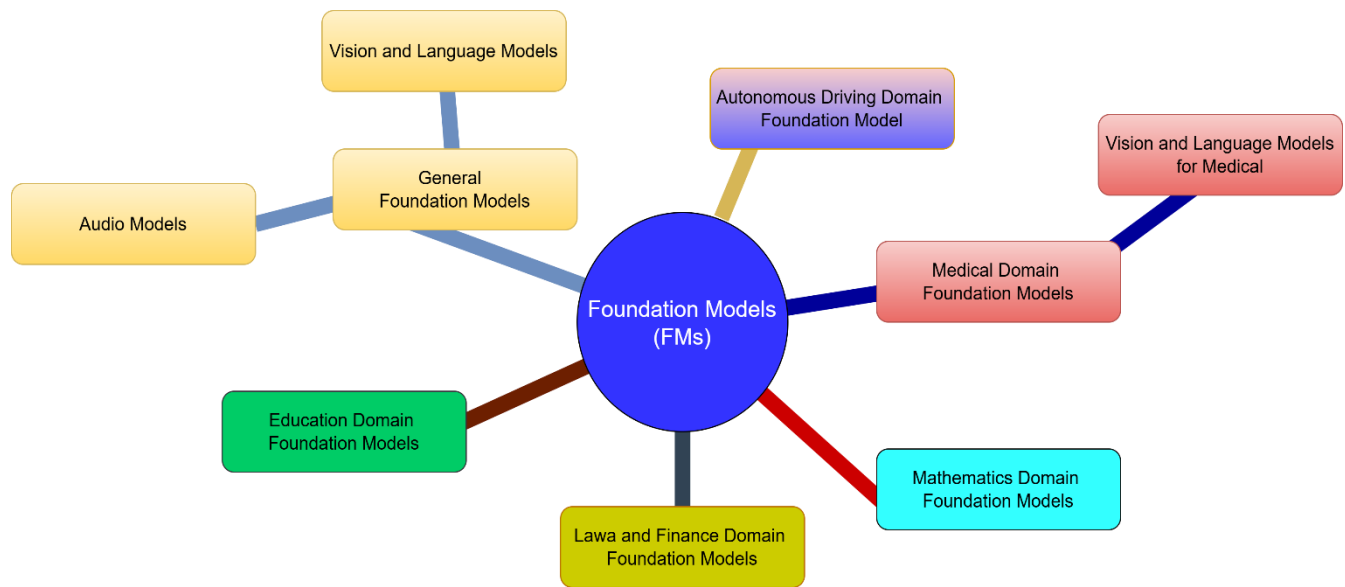


Fig. 1. An overview of Foundation models.

PRISM integrates H&E-stained histopathology images with clinical report text in its two-domain pretraining, yielding a more comprehensive and clinically relevant context. We believe what separates our approach is that the model can output clinical reports directly from slide-level embeddings, perform zero-shot cancer detection on literally millions of slides, and predict biomarkers with label-efficient training. PRISM is frequently at least as good and often better than supervised aggregator models for these converging tasks, particularly for models using linear classifiers with slide embedding. This new paradigm not only streamlines model performance, removing the inefficiencies of pooling complex tile-level information but also makes clinically actionable models, changing the landscape of cancer detection and biomarker prediction for precision medicine. Virchow's [23] model exemplifies Microsoft Research's investment in the future of medical diagnostics. Trained on an unprecedented dataset of 3.1 million whole slide images comprising 2.4 petabytes of data, Virchow comes in two advanced models of differing parameter scales. This is remarkable because it established a strong applicable recognition across multiple cancer types, a particular breakthrough of the model, with 0.95 specimen-level area under the curve for pan-cancer detection, exhibiting a great view of the power of AI in pathological research. Overall, these foundation models are a quantum leap forward in computational pathology.

BioBERT [39] is a BERT trained on PubMed's medical research papers. It can comprehend medical jargon more than normal language models. It excels in the detection of medical entities in text, concept linking, question answering, and classifying medical documents. ClinicalBERT [40] aims at physicians' notes in patient reports. BioBERT knows medical literature, but ClinicalBERT knows how doctors write about patients. It is good at retrieving useful information from clinical notes, finding medical concepts, and grouping patients based on their diseases. BlueBERT [41] was trained on medical billing codes and clinical notes. It fills the gap between official medical. Codes and how doctors describe conditions verbally. This is what makes it useful in predicting medical codes, extracting information from clinical text, and mapping codes to written

descriptions. MedBERT [42] is not one particular model but a family of BERT models trained for medical text. Each group has developed its own version by training BERT on different medical datasets. Such models perform numerous medical language tasks. GPT-4 Medical [5] is GPT-4 trained further on medical data. It is thus more appropriate for healthcare use. It can answer medical questions, aid in clinical decisions, summarize medical reports, and communicate with patients more accurately.

CheXNet [43] is a Stanford model that identifies pneumonia and other pathologies in chest X-rays. It uses a DenseNet architecture and was trained using the ChestX-ray14 dataset. It's capable of identifying some diseases as well as or better than human radiologists. MedCLIP [44] adopts the CLIP model (which associates normal images and words) and trains it with medical images and their descriptions. It helps associate medical images with medical words. It can classify medical images, generate reports, and retrieve similar medical images according to descriptions. CheXZero [45] is capable of identifying disease conditions in chest X-rays that it was not specifically trained to identify. This is a function of generalizing what it has learned about disease descriptions and connecting them to visual patterns. This allows it to identify even rare conditions. ConVIRT [46] learns through aligning medical images. It aligns one image with several views while distributing several images. This helps in learning useful patterns from unlabeled medical images. These patterns can then be used to sort, segment, and find similar images. GLORIA [47] aligns medical images with their text reports. It aligns the whole image to the whole report and individual image regions to individual sentences. This fine-grained alignment helps it generate reports, search for images given text, and answer questions about medical images. LoVT [48] connects specific areas in medical images to specific words in reports. LoVT uses transformers that can look at specific parts in images and connect them to matching text. This is helpful when you want to see exactly which area of an image corresponds to which description.

These transformers are not just a new technology set or a series of emerging tools with their fusion of advanced machine

learning techniques and large and consistent datasets. These transformers create new tools to improve diagnostic accuracy, clinical workflows, and patient outcomes. These models will further enhance personalized medicine, early disease detection, and integrative healthcare responses as research evolves. A key factor in forming these models is the scaling of the data's volume, the size of the model, and the computational power required [2, 24]. A characteristic feature of the proposed approach is that the foundation model is trained with a set of training tasks, and the model, possessing general intelligence, can solve tasks that are quite different from the training ones, using the similarities between the training tasks. This has made a great difference between this and the conventional method of developing different models from the ground for imposed problems and sub-datasets. Its wider efficiency is a matter of relevance for the readers of the journal and the subject's followers [25, 26].

C. Law and Finance Domain Foundation Models

LegalBERT [52] is a BERT model trained on legal documents. It understands legal language better than regular BERT models. It works well for reviewing legal documents, analyzing contracts, finding legal information, and sorting laws. FinBERT [53] is the financially trained variant of BERT. FinBERT is able to understand financial terms better than regular BERT. It can identify sentiment in the news, earnings statements, and social media in order to contribute to forecasting the direction of market movement. It also analyzes ESG content by determining whether text is positive, negative, or neutral. FinBERT is stronger than standard models as it is particularly trained in financial text. It uses a combination of true financial documents (such as the SEC filings) as well as synthetic data so it can become more precise. EconLLM [54] is an economic and financial analysis language model. Although little is known about specifics, it likely helps with. Predicting economic trends by scanning central bank releases and policy documents. Generating insights from the most recent financial reports or regulatory compliance. FinLLama [55] is Meta's LLaMA model customized for finance work. From Llama 2 7B and having been trained on 5 cycles of financial news, it handles sentiment, summarizes reports and classifies financial documents. It beats general models in finance work due to specialized training, and it is applied to trading, compliance, and financial analysis. DistilBERT [56] for Finance is a compact BERT model for multilingual financial sentiment analysis. Swiss multilingual financial text was used to train it, and it bypasses language barriers, requires fewer computing resources, and is efficient on low-frequency languages. Its balanced speed and accuracy make it ideal for real-time surveillance and trading platforms.

D. Mathematics Domain Foundation Models

MathBERT [57] is a BERT model that was trained on mathematics texts. It reads online maths content, textbook, and research papers. It possesses an understanding of math formulae and maths text, helping it to identify math information, understand formulae, and do mathematical reasoning. It connects the natural language with the formalism of mathematics. GPT-4 [18] Math refers to GPT-4 to perform mathematics. It is much better at math than the previous model. It can push through challenging calculus, algebra, geometry, and statistical problems. It is better because it is bigger, has design

improvements, and was trained on more math content. It is utilized by users to respond to math questions, aid in research, and instruct. Minerva [58] is Google's mathematics reasoning language model. It is a PaLM-based model that is trained on a vast amount of science and mathematics text. It solves difficult math problems, gives solutions, and explains its workings nicely. It can solve competitive math problems and math problems at the college level.

These are like models of ProofNet [59] that work with mathematical proofs. They check if proofs are correct and can create new ones. They use transformer architectures to understand the math of math statements and rules. These models help in automating math discovery and verification.

E. Autonomous Driving Domain Foundation Models

DriveLM [60] connects language models and autonomous cars. It makes cars understand scenes more effectively, as well as reason and plan. It uses Graph Visual Question Answering to connect questions and answers through logical relations. They work at two levels: object level (how cars and humans interact) and task level (connecting perception, prediction, planning, and behavior in simple terms). DriveLM performs equally well as standalone driving systems, explains decisions in simple terms, and combines vision, text, and graphs to handle complex driving situations. AutoCLIP [61] converts CLIP (a framework to connect images and text) into autonomous cars. While data is limited, AutoCLIP-like frameworks aim to improve the perception of vehicles by connecting visual information with text descriptions (e.g., connecting a camera image with "red traffic light"). AutoCLIP-like frameworks also expand reasoning by connecting driving scenarios with natural language instructions. For more information, look for papers on CLIP-based autonomous driving systems. DriveVLM [62] is a Vision Language Model implemented for self-driving cars. It translates what the car sees (roads, impediments) and determines driving directions through natural language commands like "Drive safely during heavy rain." It connects understanding vision with planning abilities to let cars drive better. DriveMLM [63] is a multimodal autonomous vehicle model that is an expert in recognizing new or unusual circumstances. It's been trained on diverse, unlabeled data to handle unusual situations like unusual weather. It can also offer explanations of driving choices in easy-to-understand language like "Slow down due to icy road," which can be employed to understand why the car takes particular choices.

F. Education Domain Foundation Models

EduBERT [64] is a BERT model fine-tuned particularly for educational text processing. It performs tasks such as learning analytics, student performance prediction, and educational content analysis using large-scale educational data to generate contextualized embeddings. The model was trained on 3.8 billion tokens of educational ASR data, and thus, it is resistant to transcription errors prevalent in online teaching settings. TAL-EduBERT [65] is a Chinese-language model developed by TAL Education Group for online learning environments. It processes educational ASR transcripts and improves tasks like question-answering and grading. The model was trained on over 20 million educational ASR transcripts to address noise like misrecognized speech, and excels in sentiment analysis and

knowledge tracing while powering TAL's internal tools for personalized learning recommendations. DistilEduBERT [66] is a lighter version of EduBERT, distilled from DistilBERT, to run faster in resource-limited settings such as mobile apps. It achieves more than 95% of the performance of the full-sized EduBERT while significantly reducing computational costs, allowing it to be deployed in day-to-day educational technology use cases.

a) Large-Scale Model Pretraining

Large-scale model pretraining is a primary approach to creating FMs from scratch. This section breaks down model size, data scale, and self-supervised learning methods as important aspects to facilitate the progress of pre-training to improve FMs.

b) The Power of Scale

The main structure of FMs is the Transformer, which has a flexible and fast training process because of its self-attention mechanism when working with a sequence of inputs. Starting in 2018, Transformer-based FMs that apply in NLP have been tuned with more massive data, thus exposing the effect of scale. For instance, the Generative Pre-trained Transformer (GPT) [30] is a large LM with 117 million parameters trained on a variety of more than 7,000 books, with 800 million words in total from BooksCorpus. In the same way, BERT [26], a new model introduced by Google, possesses 110 million parameters in the base model and 340 million in the large model, which uses BooksCorpus and English Wikipedia. Ever since, the amount of training data has skyrocketed, and the size of language models has increased from millions to billions of parameters. For example, T5 architecture is the Text-to-Text Transfer Transformer trained with 11 billion parameters and covers approximately 750 GB of the Colossal Clean Crawled Corpus (C4) dataset. GPT-3 is the third iteration of the language generation model, with 175 billion parameters, and it is trained with about 500 billion tokens of words from different sources. The models in each tier of the Large Language Model Meta AI (LLaMA) [25] vary from seven billion to 65 billion and are pre-trained on public corpora having trillions of tokens. The Pathways Language Model (PaLM) [24] has 540 billion parameters and is trained on 780 billion tokens.

c) Self-Supervised Learning

Techniques trained on lots of data need human annotation to train the model, which becomes difficult when dealing with large data sets. Common to most FM pretraining strategies, SSL is predominantly applied to learn the model without relying on labeled data. SSL research can be categorized into two primary approaches: Input reconstruction and conversion learning [26, 31]. Input reconstruction centers on the error term between the model reconstructed data and input iterate. For instance, when pre-training BERT, the Masked Language Modeling approach is applied in which the tokens within the input sequence are masked, and the model is trained to predict the masked tokens. In the same way, masked image modeling is prevalent in the case of visual representation learning, where the input image is partially masked throughout the learning process, and the model's objective is to predict the masked regions of the image. At the same time, GPT applies a different mode of thinking,

where it is trained to predict the next token in the next sequence, given the previous tokens [30, 32, 33].

On the contrary, contrastive learning tries to pull together semantically similar data points (positive samples) in the embedding space and simultaneously push apart the semantically dissimilar samples (negative samples). SimCLR [34] and MoCo-v3 [35] are models that use two random data augmentations on each training image to create positive samples and term different images as negative samples when training a vision transformer. Moreover, Contrastive Language Image Pretraining [31] leverages images and text, with images and text being positive samples and unrelated images/text being negative samples.

Lastly, there is a trend to implement input reconstruction and contrastive learning to enhance the representations. For instance, there is DINOv2 [36], which trains a student network to match the image representation to a slowly moving teacher network in the same contrastive way. The inputs here are different crops of the same image. The reconstruction of masked image patches is also challenged through the student network. The implementations of DINOv2 representations have demonstrated high competence in almost all visual evaluation indexes at an image and pixel level. In healthcare, various works have shown that self-supervised learning can fine-tune large models on large quantities of unlabeled medical data, which can work efficiently on downstream medical tasks without a lot of labeled data. For example, PubMedBERT [37], a species of a BERT model trained from text inputs without human annotation, replicates the PubMed corpora of 14 million abstracts, and this model outperforms other species in various biomedical NLP tests. Likewise, MoCo-CXR [38] was trained with a visual model on a huge number of unsupervised chest X-ray images, leading to improved image representations contrary to what would be achieved through supervised learning.

III. RISKS AND CHALLENGES

A. Privacy, Security, and Data Misuse

Foundation models are often built on massive datasets that may hold sensitive information, raising data privacy and security concerns, particularly in areas like biomedicine or law enforcement. The risk of re-identifying individuals or misusing data for predictive surveillance (e.g., predictive policing) increases ethical tensions [1].

B. Transparency, Accountability, and Trust

The black-box nature of foundation models poses a threat to transparency, where training data or decision-making are difficult to audit. This lack of explainability erodes trust, especially in critical applications like healthcare or criminal justice, where accountability is crucial [76, 77].

C. Bias, Fairness, and Cultural Sensitivity

Models learn biases from training data and generate discriminatory outputs or are culturally insensitive. These are amplified in specialized fields (e.g., medicine or law), where biased recommendations can perpetuate system inequalities or misread nuanced contexts [1, 2].

D. Societal Impact and Long-Term Accountability

The use of foundation models can destabilize labor markets, amplify disinformation, or enable malicious applications (e.g., deepfakes). Prevention of these harms requires systems for assessing long-term social impact and assigning responsibility for abuse [1].

E. Environmental and Resource Ethics

The energy used for training and running big models leads to carbon footprint and energy consumption. There is an ethical obligation to balance innovation with sustainability, particularly as models increase in size [75, 78].

F. Domain-Specific Ethical Risks

Applications in particular domains such as healthcare robots or AI for law have ethical concerns, such as patient protection, confidentiality, or professional ethics. Domain-specific tuning and rigorous testing are necessary to prevent these threats [79].

By aggregating overlapping challenges, these categories emphasize the connected ethical and societal challenges that require interdisciplinary solutions, policy frameworks, and stakeholder collaboration to build accountable AI.

IV. FUTURE OPPORTUNITIES

A. Health and Medicine

Foundation models help to diagnose patients, support medical decision-making, and document medical reports. They accelerate patient care and improve its precision. You may also train such models to create special healthcare data analysis software [80-82].

B. Automation and Efficiency

The models perform mundane tasks like summarizing documents, entering data, and running call centers. It enables humans to concentrate on more complex issues needing human intelligence [83].

C. Multimodal Capabilities

New models can process text, images, and other information at the same time. This helps in image analysis and modeling ecosystems [84].

D. Bioinformatics

The models are utilized to address biology problems in sequence analysis, protein structure, and multimedia analysis. This speeds up research in genetics and new drug discovery.

E. Finance and Education

They offer custom solutions for risk screening, fraud prevention, and personal learning platforms. This changes the business landscape of these industries [85].

F. Cross-Industry Applications

Companies can redefine AI uses for customer service, new product development, and process automation.

G. Environmental Science

Models help to study ecosystem processes such as carbon sequestration and forest structure. This helps in conserving the environment [86].

H. Language and Robotics

Enhanced language understanding and robotic capability lead to innovation in translation, voice assistants, and autonomous systems [2].

V. CONCLUSION

Foundation models are the new, most impactful innovation in AI. They have amazing capabilities, from language understanding to vision interpretation in various domains through in-depth analyses. This research explores general foundation models in various fields such as LLMs, GPT, BERT, deepseek, KIMI, and others. Our analysis points out that, though these foundation models have rapidly changed the technological perspectives and pushed the boundaries of artificial intelligence, they face great challenges in practice. Key requirements of high-quality training data, computational efficiency, and model interpretability are issues central to further improvement. However, the prospects look good, with lots of room for developing new model architectures, improving multimodal learning abilities, and creating more efficient training paradigms. In the decade to come, the ongoing advancements in the area of base models will likely not only address the gaps in today's technologies but also create a plethora of new advancements across industries. While they are precise tools for developing applications for general artificial intelligence, their strong understanding of visual and linguistic input and great flexibility are also useful for specialized medical diagnosis tools.

ACKNOWLEDGMENT

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977) and the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863).

REFERENCES

- [1] Brown, T., et al., Language models are few-shot learners. *Advances in neural information processing systems*, 2020. 33: p. 1877-1901.
- [2] Bommasani, R., et al., On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [3] Hathiya, J.J. and S. Tanwar, An exhaustive survey on security and privacy issues in Healthcare 4.0. *Computer Communications*, 2020. 153: p. 311-335.
- [4] Qiu, J., et al., Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 2023.
- [5] Nori, H., et al., Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

- [6] Christodoulou, E., et al., A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 2019. 110: p. 12-22.
- [7] Aggarwal, R., et al., Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 2021. 4(1): p. 65.
- [8] Pan, S., et al., Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [9] Zha, D., et al., Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.
- [10] Vaswani, A., Attention is All You Need. *arXiv preprint arXiv:1706.03762*, 2017.
- [11] Alfasy, S., et al., When is a foundation model a foundation model. *arXiv preprint arXiv:2309.11510*, 2023.
- [12] Johnson, J.M. and T.M. Khoshgoftaar, Data-centric ai for healthcare fraud detection. *SN Computer Science*, 2023. 4(4): p. 389.
- [13] Singh, P., Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 2023. 6(3): p. 144-157.
- [14] Hunter, B., S. Hindocha, and R.W. Lee, The role of artificial intelligence in early cancer diagnosis. *Cancers*, 2022. 14(6): p. 1524.
- [15] Chen, W., et al., Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular Therapy-Nucleic Acids*, 2023. 31: p. 691-702.
- [16] Rao, A., et al., Assessing the utility of ChatGPT throughout the entire clinical workflow. *MedRxiv*, 2023: p. 2023.02. 21.23285886.
- [17] Ozmen Garibay, O., et al., Six human-centered artificial intelligence grand challenges. *International Journal of Human-Computer Interaction*, 2023. 39(3): p. 391-437.
- [18] OpenAI, GPT Series: GPT-3, GPT-4, and GPT-5 (in development) from OpenAI. p. OpenAI is an AI research and deployment company. Our mission is to ensure that artificial general intelligence benefits all of humanity. <https://openai.com/news/>
- [19] Devlin, J., et al. Bert: Pretraining of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.
- [20] Raffel, C., et al., Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020. 21(140): p. 1-67.
- [21] Liu, Y., et al., Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [22] Yang, Z., et al., Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019. 32.
- [23] DeepSeek. [cited 2025 03.01]; DeepSeekR1 is a large language model developed by DeepSeek AI. Available from: <https://www.deepseek.com/>.
- [24] Qwen-VL. [cited 2025 03.01]; Qwen-VL (Qwen Large Vision Language Model) is the visual multimodal version of the large model series, Qwen (abbr. Tongyi Qianwen), proposed by Alibaba Cloud.]. Available from: <https://modelscope.cn/models/qwen/Qwen-VL-Chat/summary>.
- [25] AI, K. p. Kimi AI is an advanced AI model built by Moonshot AI in Beijing. <https://dev.to/arbisoftcompany/whats-the-hype-about-kimi-ai-141>.
- [26] AI, M. [cited 2025 03.01]; Mistral AI is a French company developing large language models with a focus on efficiency, performance, and open access.]. Available from: <https://mistral.ai/en>.
- [27] Radford, A., et al. Learning transferable visual models from natural language supervision. in *International conference on machine learning*. 2021. PMLR.
- [28] Dosovitskiy, A., et al., An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [29] Liu, Z., et al. Swin transformer: Hierarchical vision transformer using shifted windows. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
- [30] DALL·E. [cited 2025 03.01]; DALL·E is a 12-billion parameter version of GPT-3(opens in a new window) trained to generate images from text descriptions, using a dataset of text-image pairs.]. Available from: <https://openai.com/index/dall-e/>.
- [31] Baevski, A., et al., wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020. 33: p. 12449-12460.
- [32] Hsu, W.-N., et al., Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 2021. 29: p. 3451-3460.
- [33] Chen, R.J., et al., Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 30(3): p. 850-862.
- [34] Lu, M.Y., et al., A visual-language foundation model for computational pathology. *Nature Medicine*, 2024. 30(3): p. 863-874.
- [35] Yun, J., et al., EXAONEPath 1.0 Patch-level Foundation Model for Pathology. *arXiv preprint arXiv:2408.00380*, 2024.
- [36] Kang, M., et al. Benchmarking self-supervised learning on diverse pathology datasets. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- [37] Shaikovski, G., et al., PRISM: A Multimodal Generative Foundation Model for Slide-Level Histopathology. *arXiv preprint arXiv:2405.10254*, 2024.
- [38] Vorontsov, E., et al., Virchow: A Million-Slide Digital Pathology Foundation Model. *arXiv 2023. arXiv preprint arXiv:2309.07778*.
- [39] Lee, J., et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020. 36(4): p. 1234-1240.
- [40] Huang, K., J. Altosaar, and R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.
- [41] Peng, Y., S. Yan, and Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
- [42] Rasmy, L., et al., Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 2021. 4(1): p. 86.
- [43] Rajpurkar, P., et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [44] Wang, Z., et al. Medclip: Contrastive learning from unpaired medical images and text. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. 2022.
- [45] Lai, H., et al. Carzero: Cross-attention alignment for radiology zero-shot classification. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [46] Zhang, Y., et al. Contrastive learning of medical visual representations from paired images and text. in *Machine learning for healthcare conference*. 2022. PMLR.
- [47] Huang, S.-C., et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
- [48] Müller, P., et al. Joint learning of localized representations from medical images and reports. in *European conference on computer vision*. 2022. Springer.
- [49] Chowdhery, A., et al., Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023. 24(240): p. 1-113.
- [50] Touvron, H., et al., Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [51] Devlin, J., Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [52] Chalkidis, I., et al., LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.
- [53] Araci, D., Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
- [54] Guo, Y. and Y. Yang, EconNLI: Evaluating Large Language Models on Economics Reasoning. *arXiv preprint arXiv:2407.01212*, 2024.

- [55] Iacovides, G., et al. FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading. in Proceedings of the 5th ACM International Conference on AI in Finance. 2024.
- [56] Meta, A., Introducing LLaMA: A foundational, 65-billion-parameter large language model. Meta AI, 2023.
- [57] Peng, S., et al., Mathbert: A pre-trained model for mathematical formula understanding. arXiv preprint arXiv:2105.00377, 2021.
- [58] Dyer, E. and G. Gur-Ari, Minerva: Solving quantitative reasoning problems with language models. June, 2022. 30: p. 2022.
- [59] Azerbayev, Z., et al., Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. arXiv preprint arXiv:2302.12433, 2023.
- [60] Sima, C., et al., Driving with graph visual question answering. arXiv 2023. arXiv preprint arXiv:2312.14150, 2023.
- [61] Lin, S., et al. Common diffusion noise schedules and sample steps are flawed. in Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024.
- [62] Tian, X., et al., Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289, 2024.
- [63] DriveMLM. p. 2024 Autonomous Driving (Multimodal) Big Model Overview: From DriveGPT4, DriveMLM to DriveVLM, DriveMM. https://blog.csdn.net/v_JULY_v/article/details/141113628.
- [64] Clavié, B. and K. Gal, Edubert: Pretrained deep language models for learning analytics. arXiv preprint arXiv:1912.00690, 2019.
- [65] TAL-EduBERT. Available from: <https://ai.100tal.com/?spm=2b75ac3d.2ef5001f.0.0.3d4dc921IthxiB>.
- [66] DistilEduBERT. A lightweight version of EduBERT, distilled from DistilBERT for faster inference in resource-constrained settings. [cited 2025 03.01]; Available from: <https://huggingface.co/bclavie/distiledubert?spm=2b75ac3d.2ef5001f.0.0.3d4dc921IthxiB>.
- [67] Radford, A., et al., Improving language understanding by generative pretraining. 2018.
- [68] Xie, Z., et al. Simmim: A simple framework for masked image modeling. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [69] He, K., et al. Masked autoencoders are scalable vision learners. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [70] Chen, T., et al. A simple framework for contrastive learning of visual representations. in International conference on machine learning. 2020. PMLR.
- [71] Chen, X., S. Xie, and K. He. An empirical study of training self-supervised vision transformers. in Proceedings of the IEEE/CVF international conference on computer vision. 2021.
- [72] Oquab, M., et al., Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.
- [73] Gu, Y., et al., Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 2021. 3(1): p. 1-23.
- [74] Sowrirajan, H., et al. Moco pretraining improves representation and transferability of chest x-ray models. in Medical Imaging with Deep Learning. 2021. PMLR.
- [75] Domínguez Hernández, A., et al. Mapping the individual, social and biospheric impacts of Foundation Models. in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024.
- [76] Challenges of Evaluating and Understanding Foundation models. [cited 2025 03.01]; Available from: <https://dev.to/azure/challenges-of-evaluating-and-understanding-foundation-models-320g>.
- [77] AI and Ethics: Navigating the Challenges and Opportunities. [cited 2025 03.01]; Available from: <https://www.gisma.com/blog/ai-and-ethics-navigating-the-challenges-and-opportunities>.
- [78] Xiao, X., et al., Robot learning in the era of foundation models: A survey. arXiv preprint arXiv:2311.14379, 2023.
- [79] Stahl, B.C. and M. Coeckelbergh, Ethics of healthcare robotics: Towards responsible research and innovation. Robotics and Autonomous Systems, 2016. 86: p. 152-161.
- [80] AWS. What are Foundation Models? [cited 2025 03.01]; Available from: <https://aws.amazon.com/what-is/foundation-models/>.
- [81] Wójcik, M.A. Foundation models in healthcare: opportunities, biases and regulatory prospects in Europe. in International Conference on Electronic Government and the Information Systems Perspective. 2022. Springer.
- [82] Moor, M., et al., Foundation models for generalist medical artificial intelligence. Nature, 2023. 616(7956): p. 259-265.
- [83] V7. Foundation Models: The Benefits, Risks, and Applications. [cited 2025 03.01]; Available from: <https://www.v7labs.com/blog/foundation-models-guide>.
- [84] The Digital Platform Regulators Forum (DP-REG). [cited 2025 03.01]; Available from: <https://dp-reg.gov.au/working-paper-3-examination-technology-multimodal-foundation-models>.
- [85] Küchemann, S., et al., Are large multimodal foundation models all we need? On opportunities and challenges of these models in education. EdArXiv, 2024.
- [86] Tan, C., et al., On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. arXiv preprint arXiv:2312.17016, 2023.

Ali Hussain received his BS degree in Computer Science from Government College University Faisalabad (GCUF), Pakistan, in 2019, and his MS degree in Digital Anti-aging healthcare from Inje University, South Korea, in 2022. He is a PhD student and research assistant at the Department of Digital Anti-aging Healthcare at Inje University, South Korea. His research interests include data science, machine learning, deep learning, computer vision, image classification, medical images, data mining, blockchain, Metaverse, and the Self-Supervised Learning, Foundation model.

Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chungbuk National University, the Republic of Korea. Furthermore, he earned Ph.D. degree from Inje University South Korea, majoring in Artificial Intelligence in healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Self-supervised learning, Computer vision, and medical imaging.

Umm E Farwa received her BS degree in Information Technology from the University of Sargodha, Pakistan., she is now an MS candidate at Inje University South Korea, majoring in Artificial Intelligence in healthcare. Her research interests include artificial intelligence, data science, big data, machine learning, deep learning, self-supervised learning, Computer vision, Foundation Models, and medical imaging.

Md Ariful Islam Mozumder received his BSc in Computer Science & Engineering from the World University of Bangladesh and an MS degree in Artificial Intelligence from the Inje University South Korea in 2022. Currently, he is pursuing Ph.D. in the Institute of Digital Anti-aging at Inje University. His research interests align with machine learning, deep learning, medical image processing, digital pathology images, bio-signal processing, and computer vision.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in SoGang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision