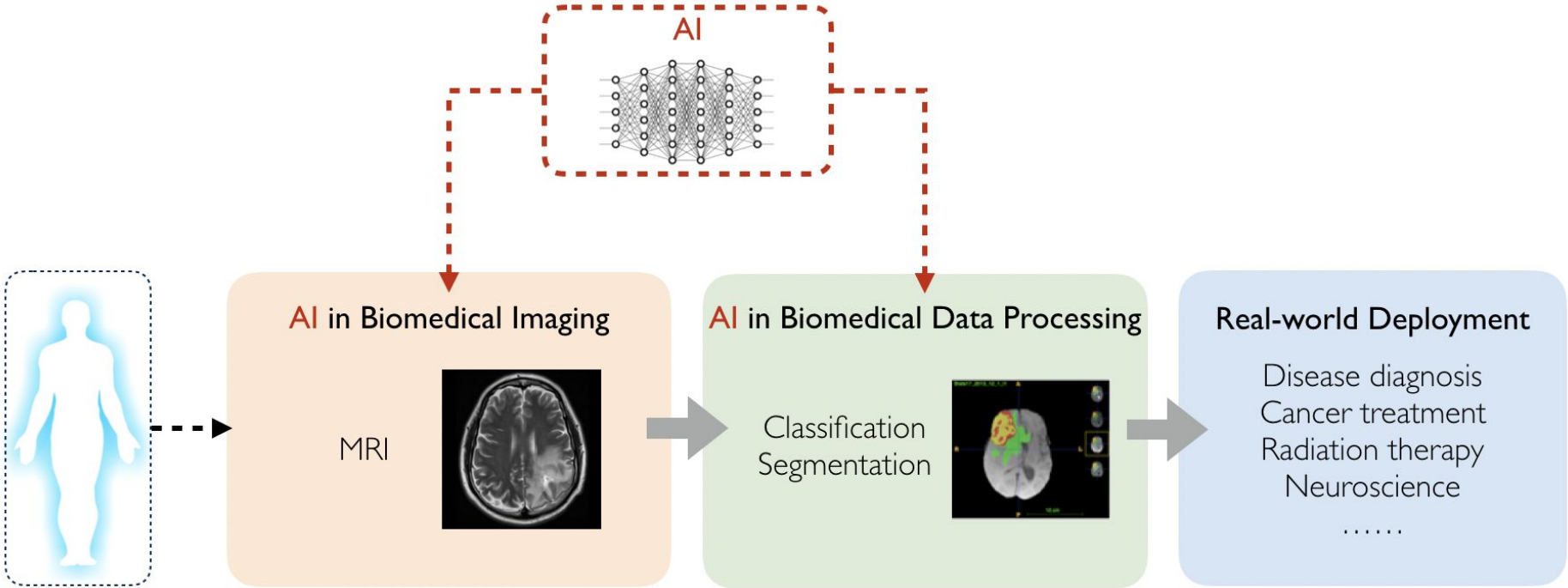
Lecture 6:
Multimodal Foundation Models

In this class:

- Part I: AI in biomedical imaging
- Part II: AI in biomedical data processing



Today's agenda

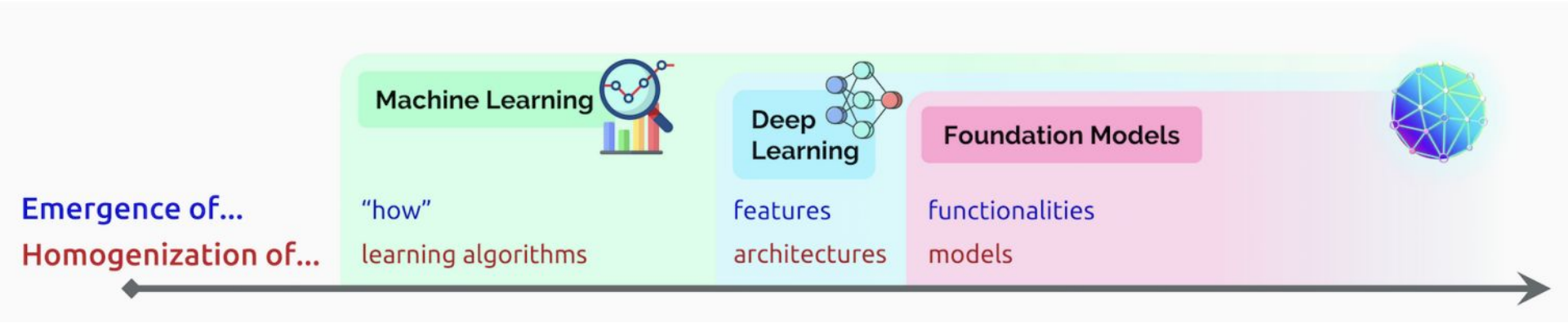
- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Today's agenda

- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Foundation model

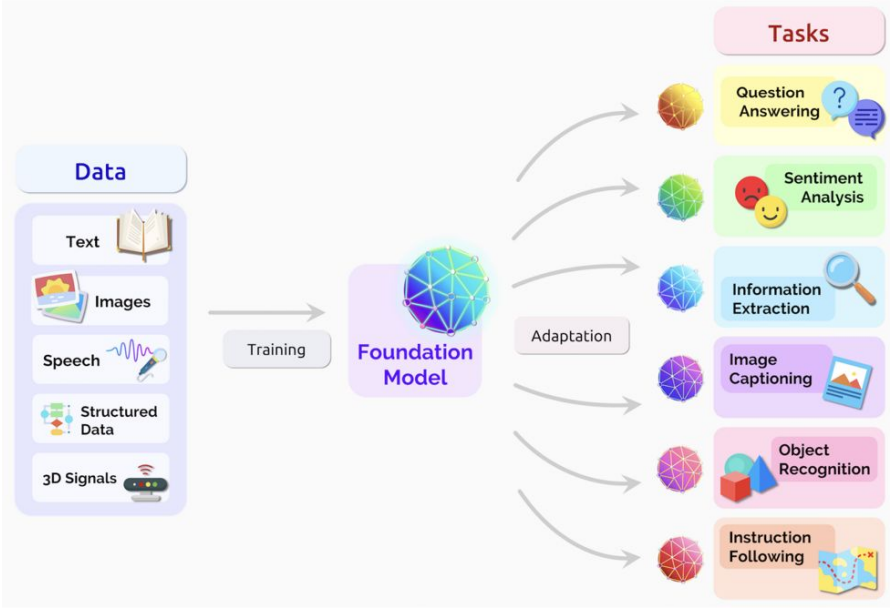
- A new paradigm for building AI systems



Bommasani, et al., On the Opportunities and Risks of Foundation Models, arXiv 2022.

Foundation model

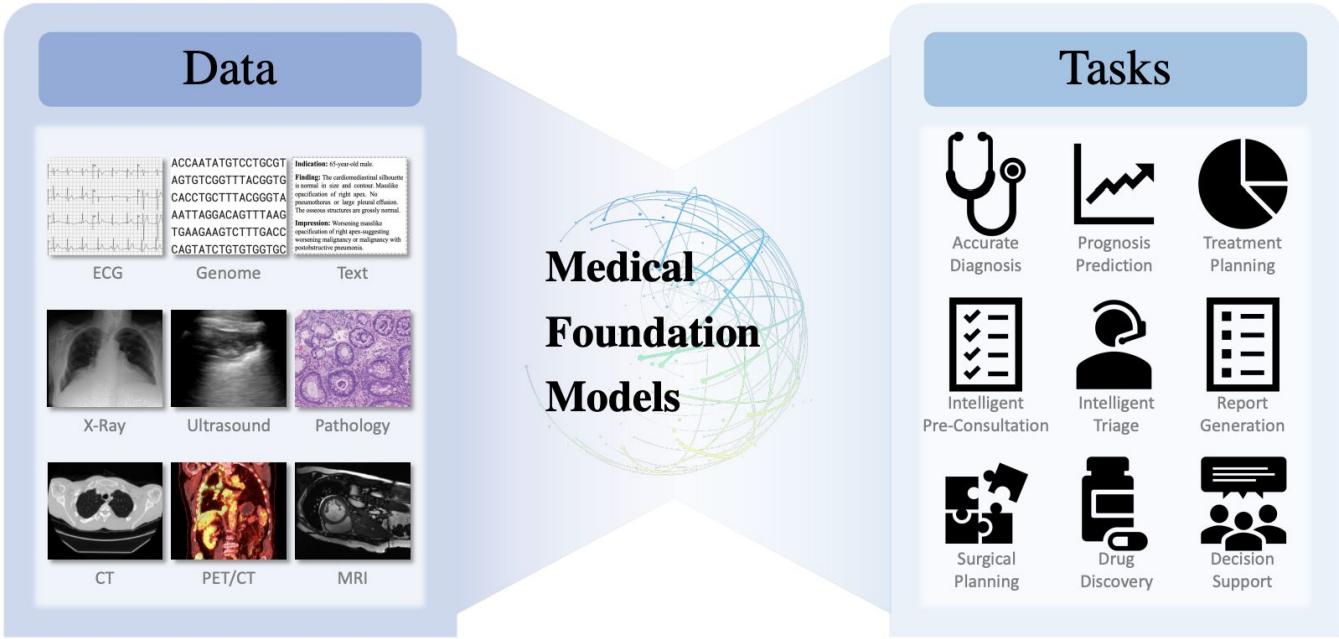
- A new paradigm for building AI systems
  - Train one model on a huge amount of multimodal data using self-supervision at scale
  - Can adapt to a wide range of downstream tasks



Bommasani, et al., On the Opportunities and Risks of Foundation Models, arXiv 2022.

Medical foundation model

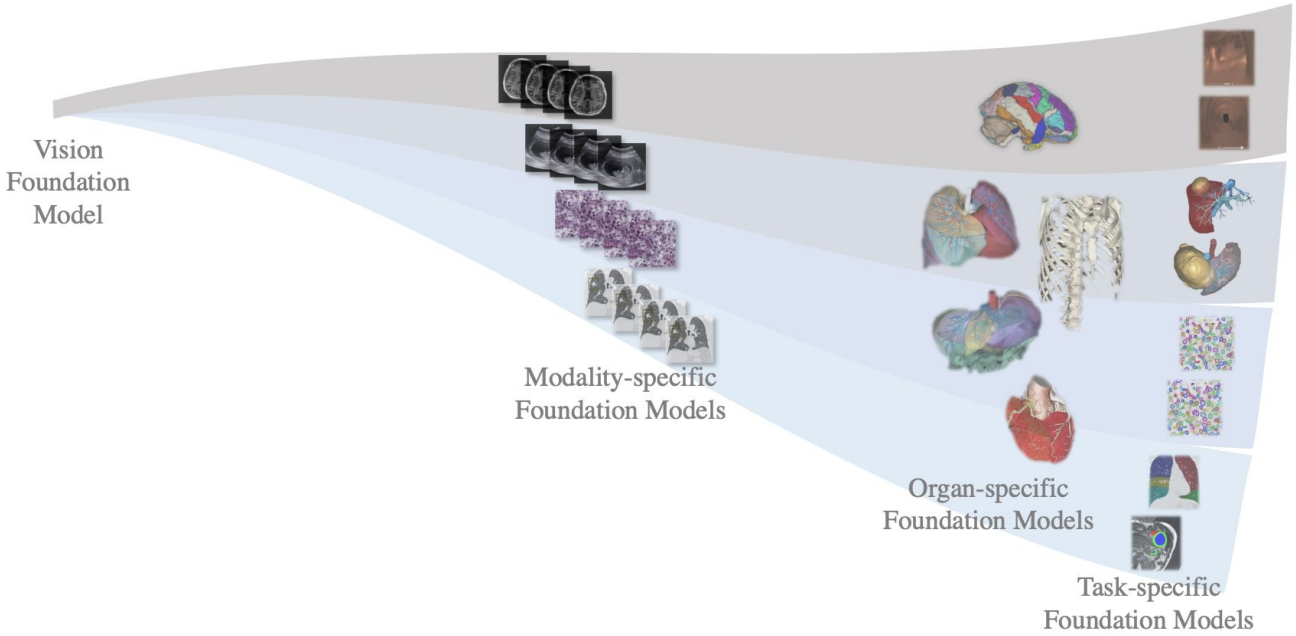
- Models trained on multiple medical data modalities can enable comprehensive clinical solutions



Zhang, et al., On the Challenges and Perspectives of Foundation Models for Medical Image Analysis, arXiv 2023.

Medical foundation model

- Modality-specific and organ-specific foundation models



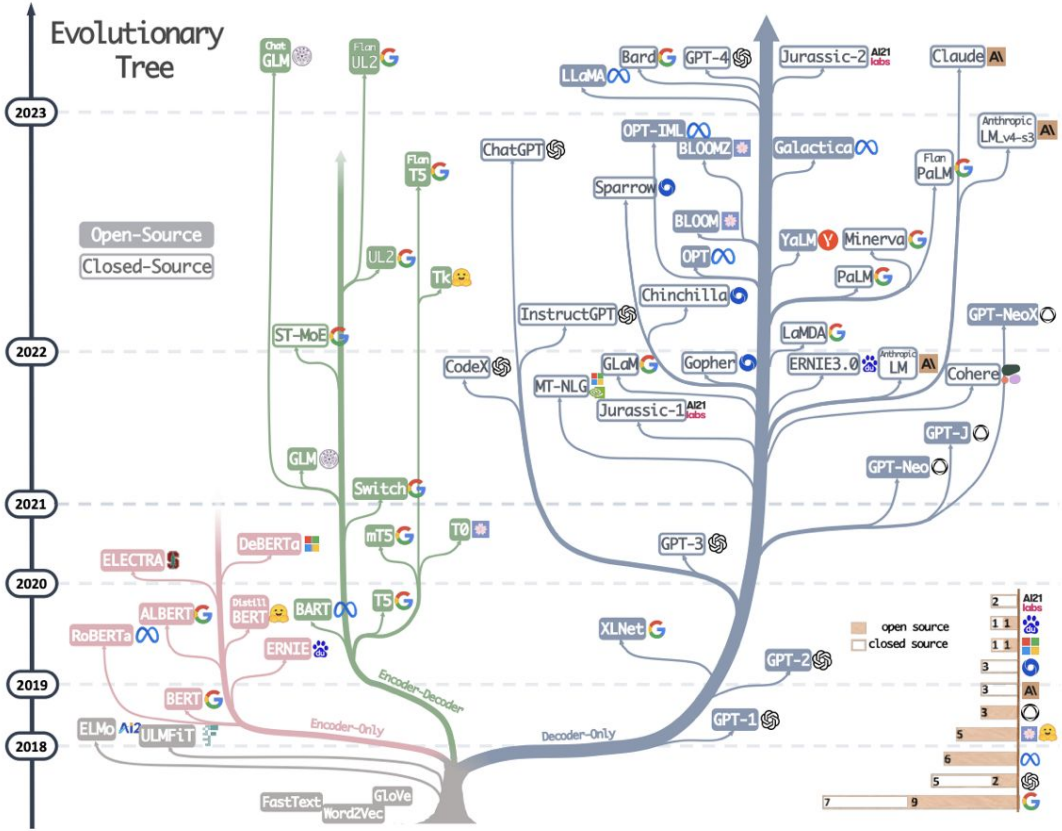
Zhang, et al., On the Challenges and Perspectives of Foundation Models for Medical Image Analysis, arXiv 2023.

Today's agenda

- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Language foundation model

- Large Language Models (LLMs)
  - Zero-shot generalization ability



Yang, et al., Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond, arXiv 2023.

Language foundation model

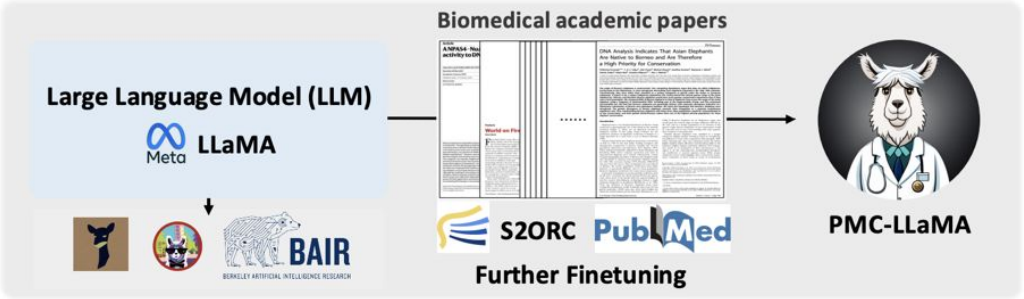
- Large Language Models (LLMs)
  - BERT-style language models: encoder-decoder or encoder-only
  - GPT-style language models: decoder-only

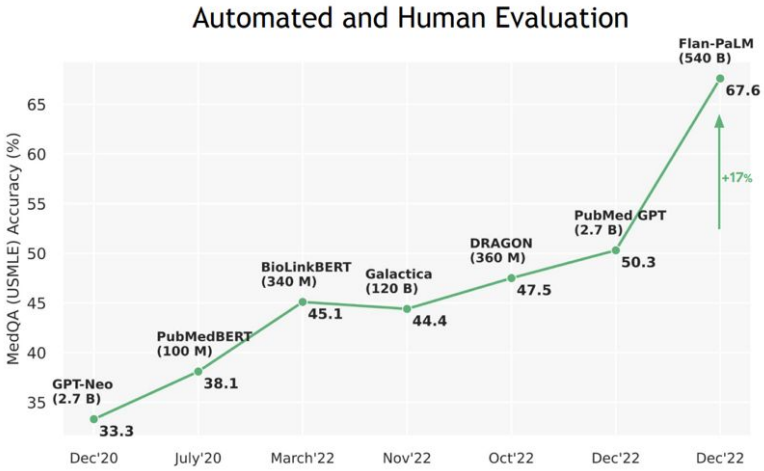
	Characteristic		LLMs
Encoder-Decoder or Encoder-only (BERT-style)	Training: Model type: Pretrain task:	Masked Language Models Discriminative Predict masked words	ELMo [80], BERT [28], RoBERTa [65], DistilBERT [90], BioBERT [57], XLM [54], Xlnet [119], ALBERT [55], ELECTRA [24], T5 [84], GLM [123], XLM-E [20], ST-MoE [133], AlexaTM [95]
Decoder-only (GPT-style)	Training Model type: Pretrain task:	Autoregressive Language Models Generative Predict next word	GPT-3 [16], OPT [126]. PaLM [22], BLOOM [92], MT-NLG [93], GLaM [32],Gopher [83], chinchilla [41], LaMDA [102], GPT-J [107], LLaMA [103], GPT-4 [76], BloombergGPT [117]

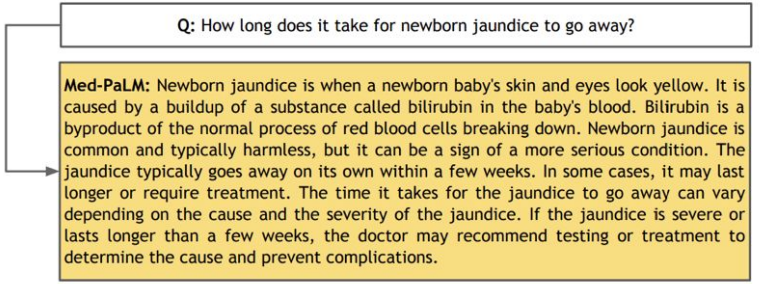
Yang, et al., Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond, arXiv 2023.

Medical language foundation model

- Medical LLMs
  - Train from scratch on medical language
    - BioGPT
    - BioMedLM (2.7-billion)
  - Fine-tune from LLMs of natural language
    - PMC-LLaMA (7-billion)
    - MedPaLM (540-billion)







Med-PaLM performs encouragingly on consumer medical question answering

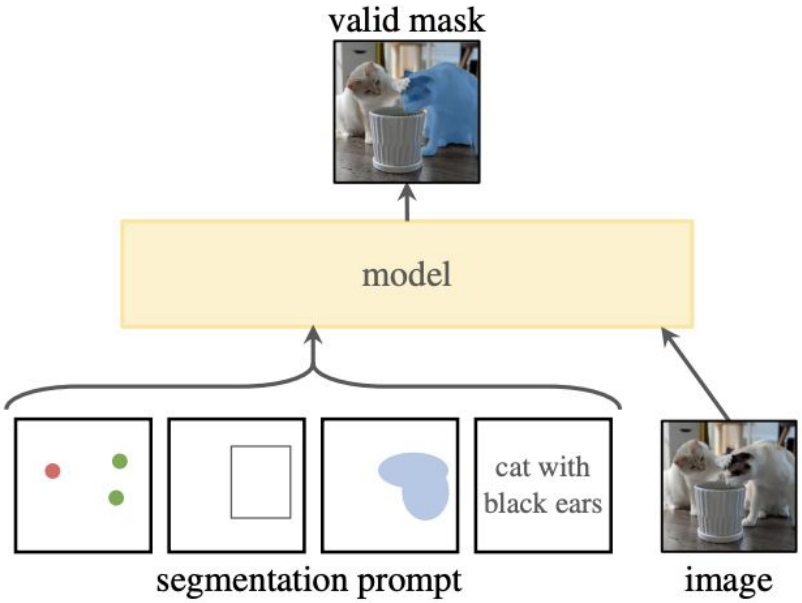
Luo, et al., BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics 2022.
Singhal, et al., Large language models encode clinical knowledge, arXiv 2022.
Wu, et al., PMC-LLaMA: Further finetuning LLaMA on medical papers, arXiv 2023.

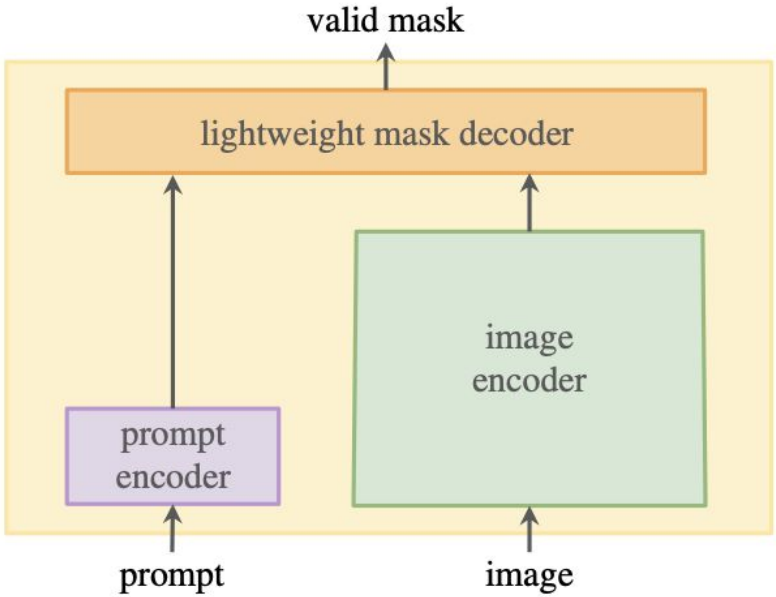
Today's agenda

- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Vision foundation model

- Segment Anything Model (SAM) for image segmentation
  - Enable zero-shot transfer to a range of tasks via prompt engineering





Kirillov et al. Segment Anything. arXiv 2023.

Medical vision foundation model

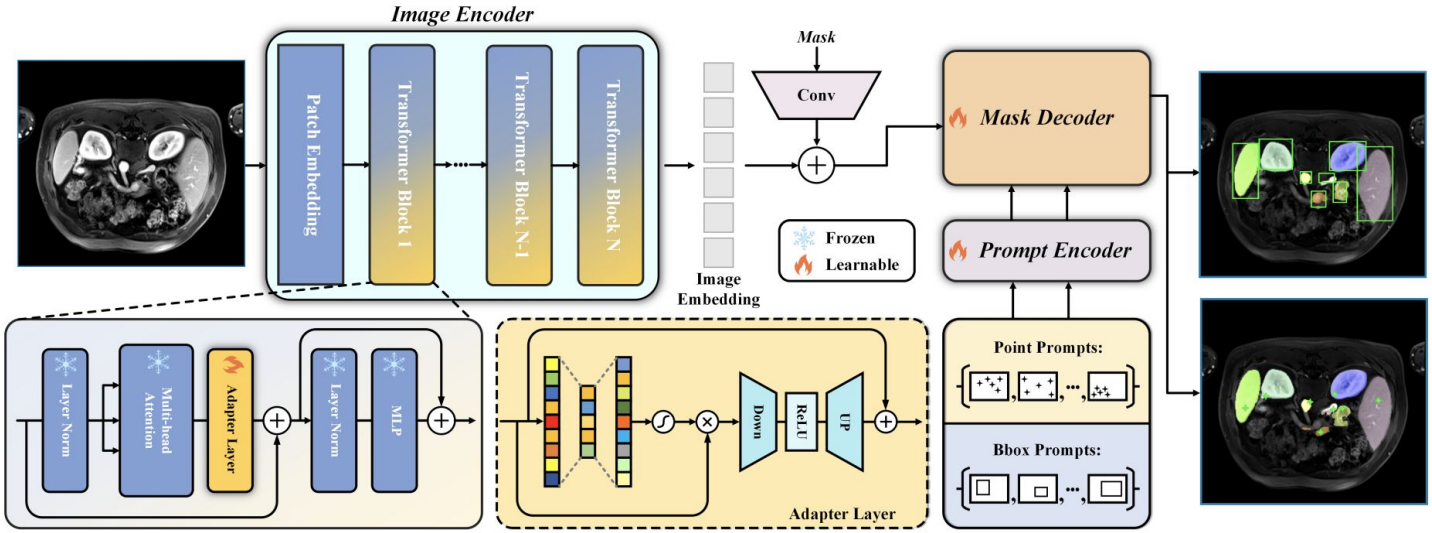
- Fine-tune SAM for medical images
  - Support multiple prompts

Model	Dataset (size)	Encoder (or Adapter)	Prompts mode			Decoder
			Point	*Bbox*	*Mask*	
SAM-U [16]	6000 masks	✘	✘	✔	✘	✘
SAMed [17]	3779 masks	✔	✘	✘	✘	✔
AutoSAM [18]	ACDC [19]	✘	✘	✘	✘	✔
MedSAM [20]	~1.1M masks	✘	✘	✔	✘	✔
MSA [21]	5 datasets	✔	✔	✘	✘	✔
SAM-Med2D (Ours)	~19.7M masks	✔	✔	✔	✔	✔

Cheng et al. SAM-Med2D. arXiv 2023.

Medical vision foundation model

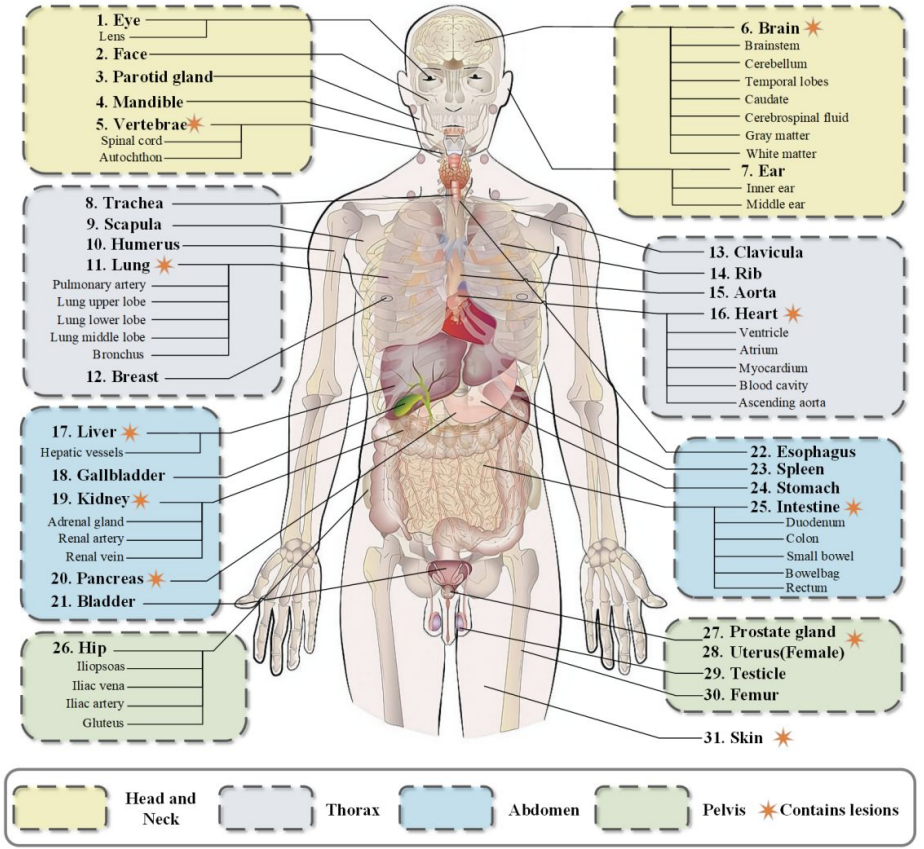
- Fine-tune SAM for medical images
  - Incorporate learnable adapter layers in each Transformer block to acquire domain-specific knowledge in the medical field
  - Fine-tune the prompt encoder using point, Bbox, and mask information, and update mask decoder



Cheng et al. SAM-Med2D. arXiv 2023.

Medical vision foundation model

- Fine-tune SAM for medical images
  - Datasets used cover 31 major organs



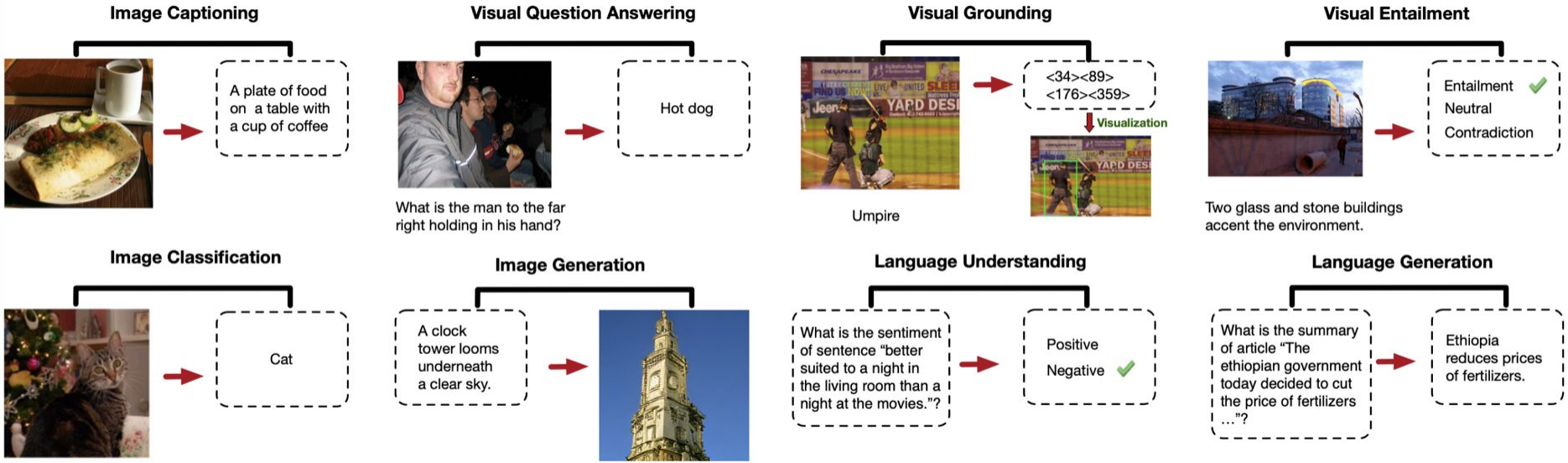
Cheng et al. SAM-Med2D. arXiv 2023.

Today's agenda

- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Multimodal foundation model

- Multimodal pre-training paradigm for general vision-language understanding
  - Learn transferable cross-modal representations from various self-supervised tasks
  - Generalize to advance various downstream tasks through fine-tuning



Wang, et al., OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, ICML 2022.

Multimodal foundation model

- Multimodal pre-training paradigm for general vision-language understanding
  - Learn transferable cross-modal representations from various self-supervised tasks
  - Generalize to advance various downstream tasks through fine-tuning

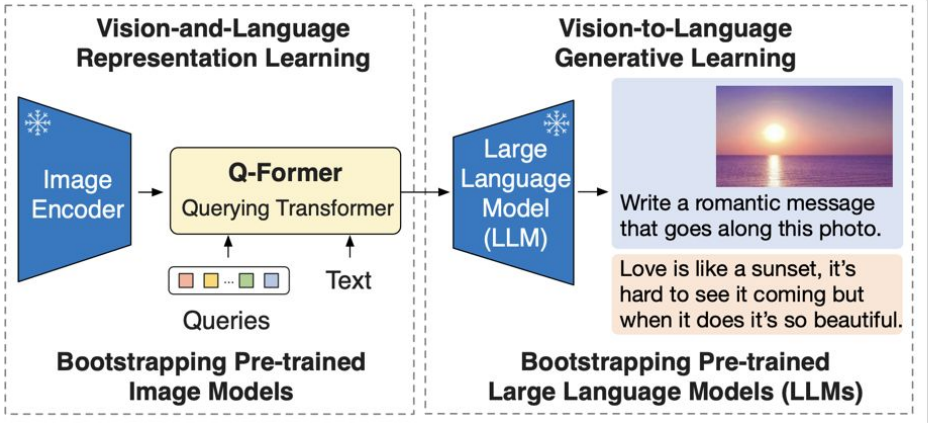
	Computer Vision				Natural Language Processing			Image-Text				Video-Text		
Method	Image Cls.	Video Cls.	Det.	Seg.	Text Cls.	QA	Summarization	Retrieval	QA	Captioning	VG	Retrieval	QA	Captioning
BEiT-3	✓		✓	✓				✓	✓	✓				
EVA	✓	✓	✓	✓										
CLIP	✓							✓				✓		
ALBEF								✓	✓		✓			
BLIP								✓	✓	✓		✓	✓	
VATT	✓	✓										✓		
Florence	✓	✓	✓					✓	✓			✓		
CoCa	✓	✓						✓	✓	✓		✓		
VideoCoCa		✓												
Flamingo		✓						✓	✓	✓			✓	✓
GIT2	✓				✓				✓	✓			✓	✓
FLAVA	✓				✓	✓	✓		✓					
OFA	✓				✓	✓	✓	✓	✓	✓	✓			
OmniVL	✓	✓						✓	✓	✓		✓	✓	✓
mPLUG 2.0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Xu, et al., mPLUG-2: A Modularized multi-modal foundation model across text, image and video, arXiv 2023.

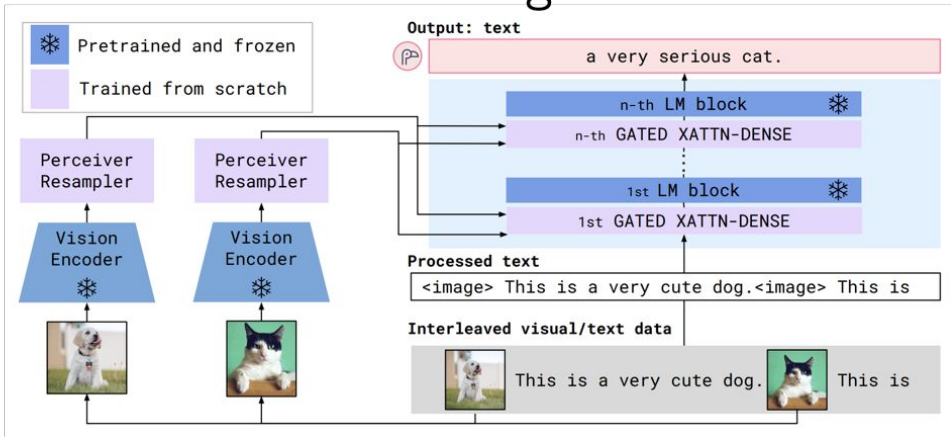
Multimodal foundation model based on LLM

- Enhance vision-language understanding using LLM
  - Learn connection module between visual encoder and LLMs
  - BLIP-2, Flamingo, FROMAGe, MiniGPT-4 …

BLIP-2



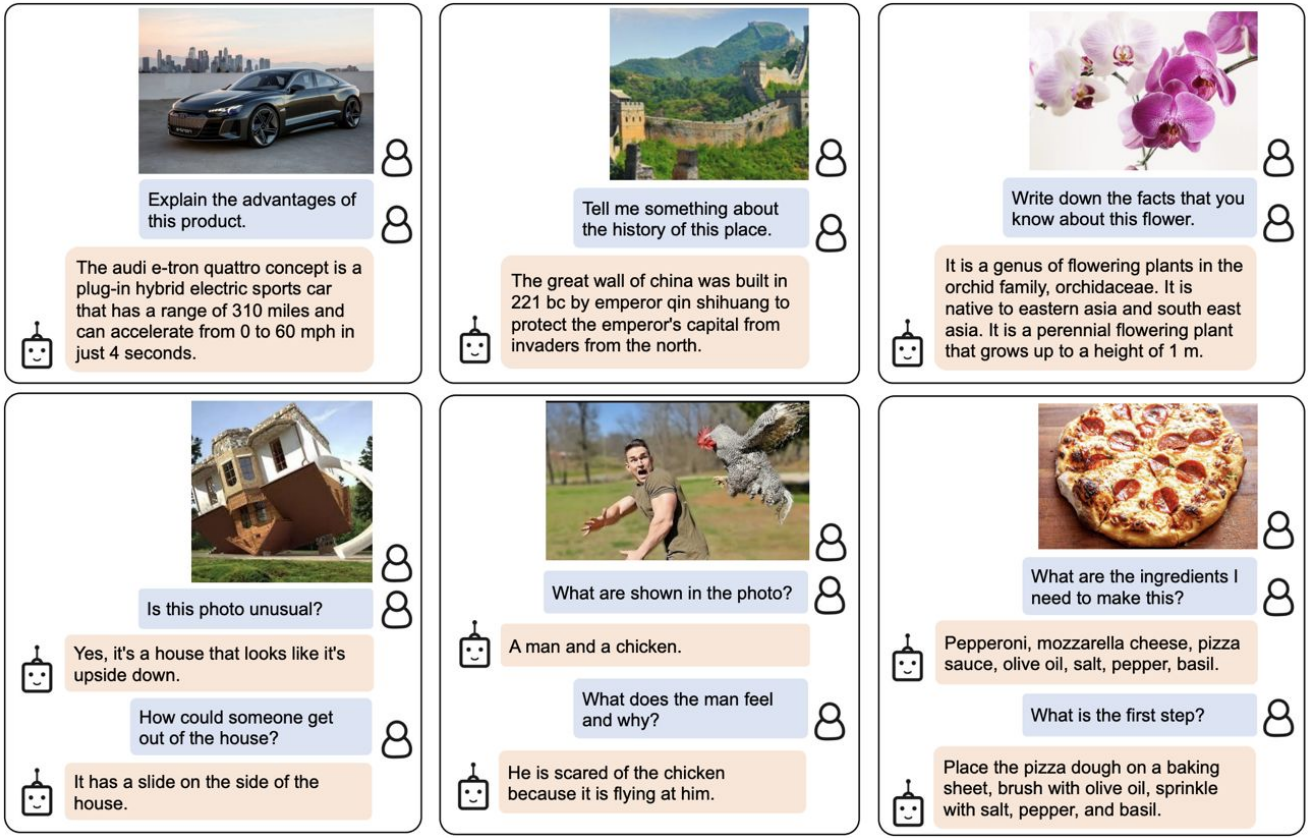
Flamingo



Alayrac, et al., Flamingo: a Visual Language Model for Few-Shot Learning, NeurIPS 2022.
Li, et al., BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv 2023.

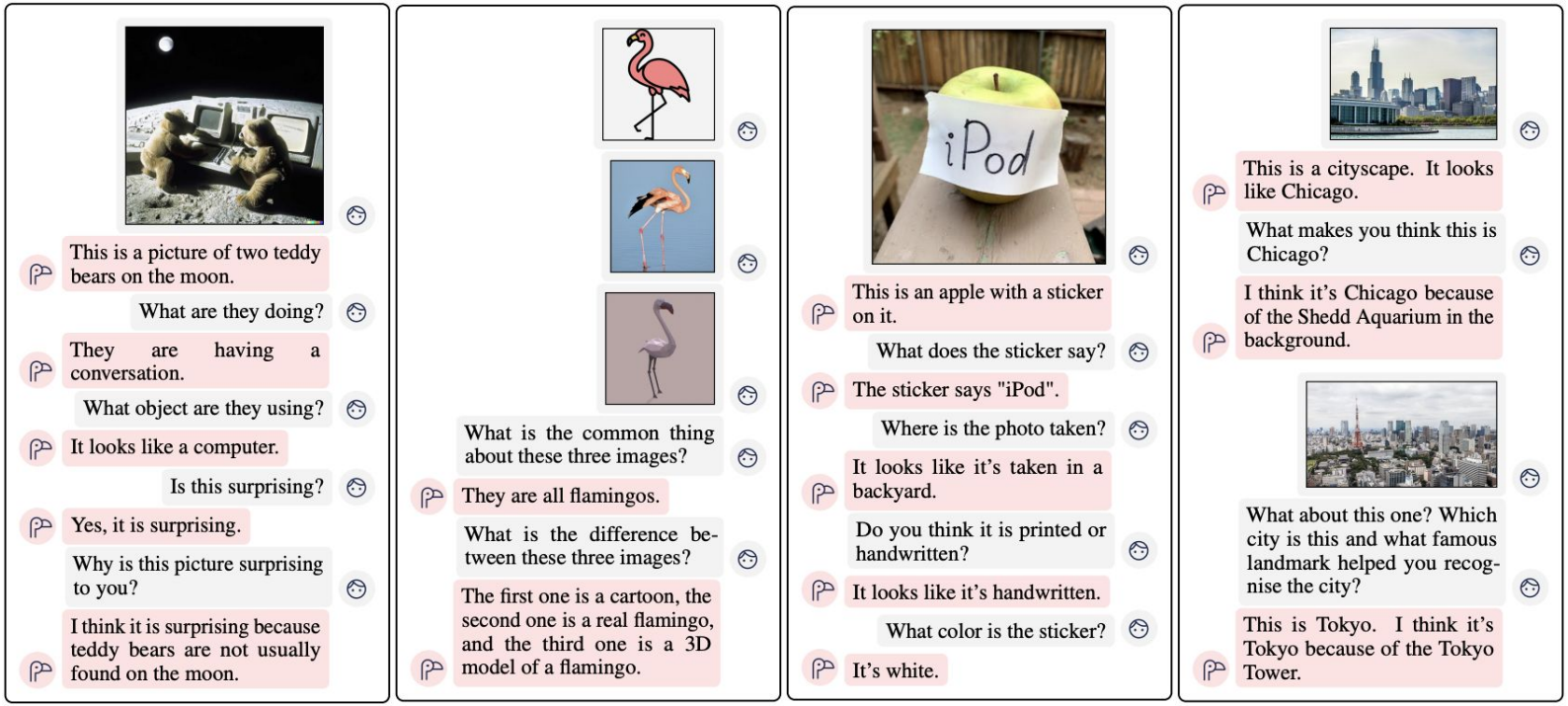
Multimodal foundation model based on LLM

- BLIP-2:
  - Instructed zero-shot image-to-text generat



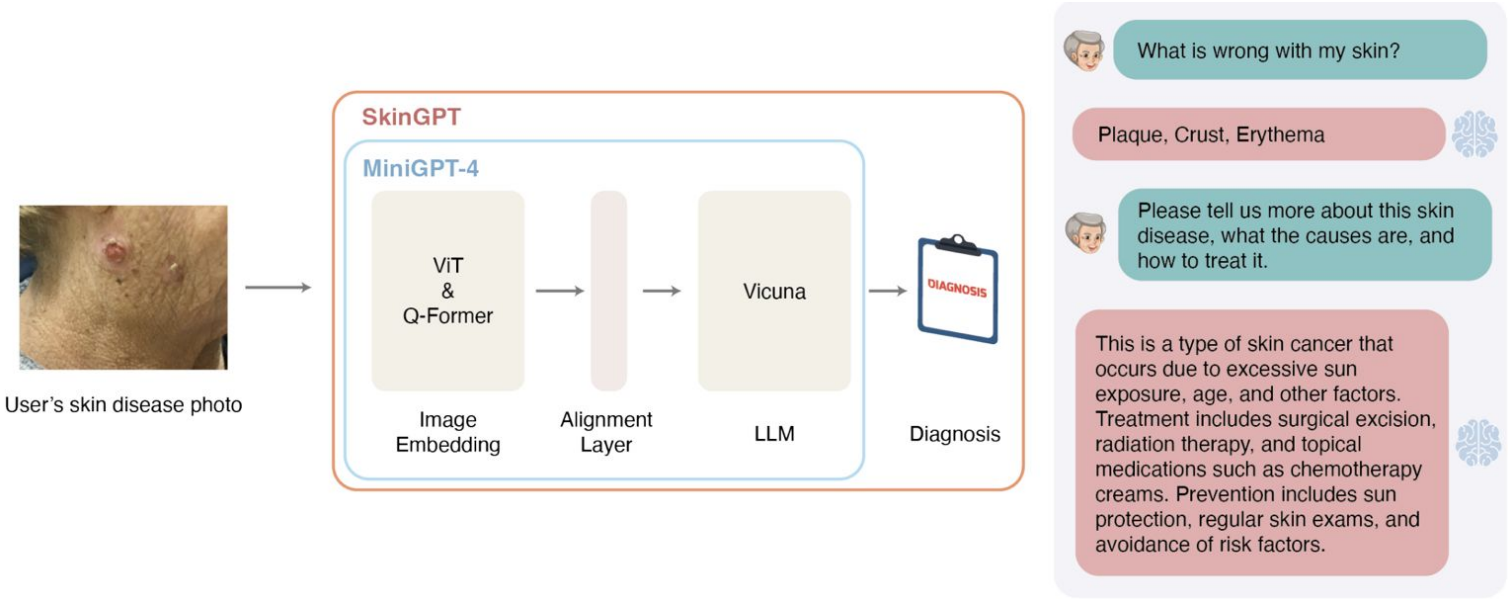
Multimodal foundation model based on LLM

- Flamingo:
  - Multi-image visual dialogue



Medical multimodal foundation model based on LLM

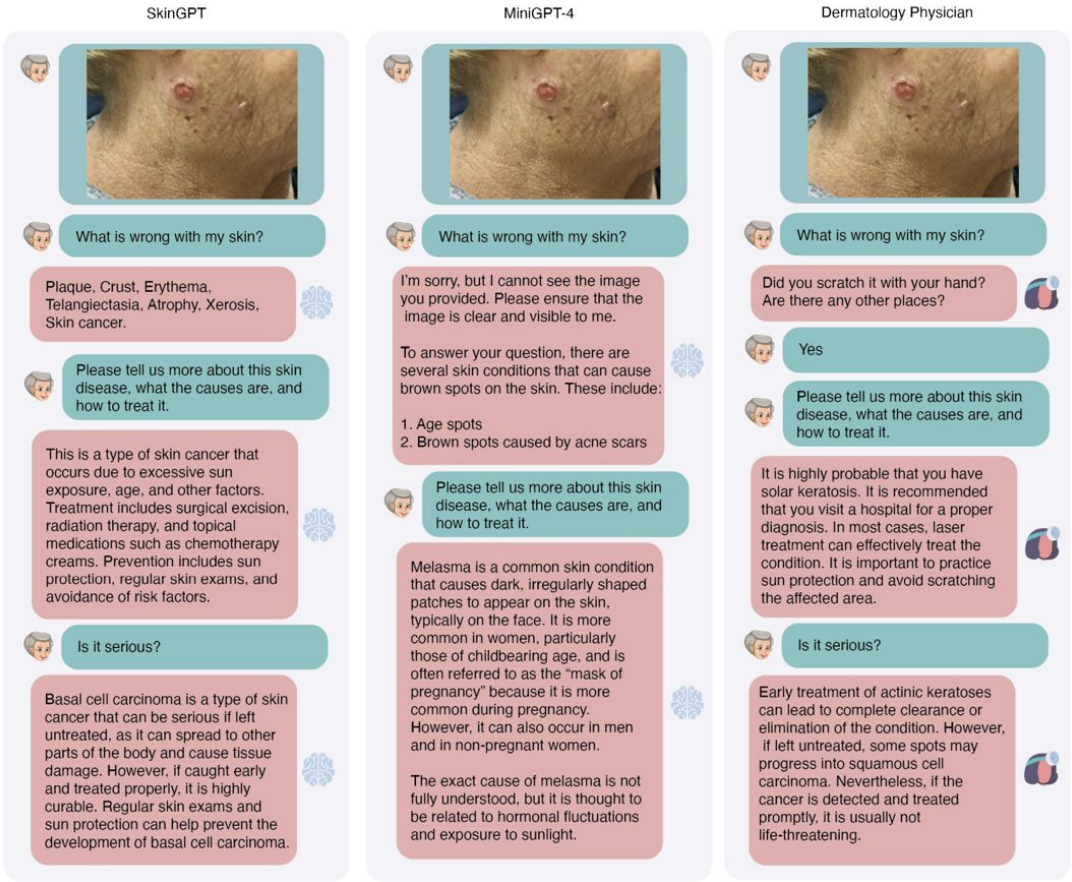
- SkinGPT-4
  - Fine-tune from MiniGPT-4 with skin disease images
  - Enable interactive dialogue with multimodal inputs



Zhou, et al., SkinGPT-4: An Interactive Dermatology Diagnostic System with Visual Large Language Model, arXiv 2023.

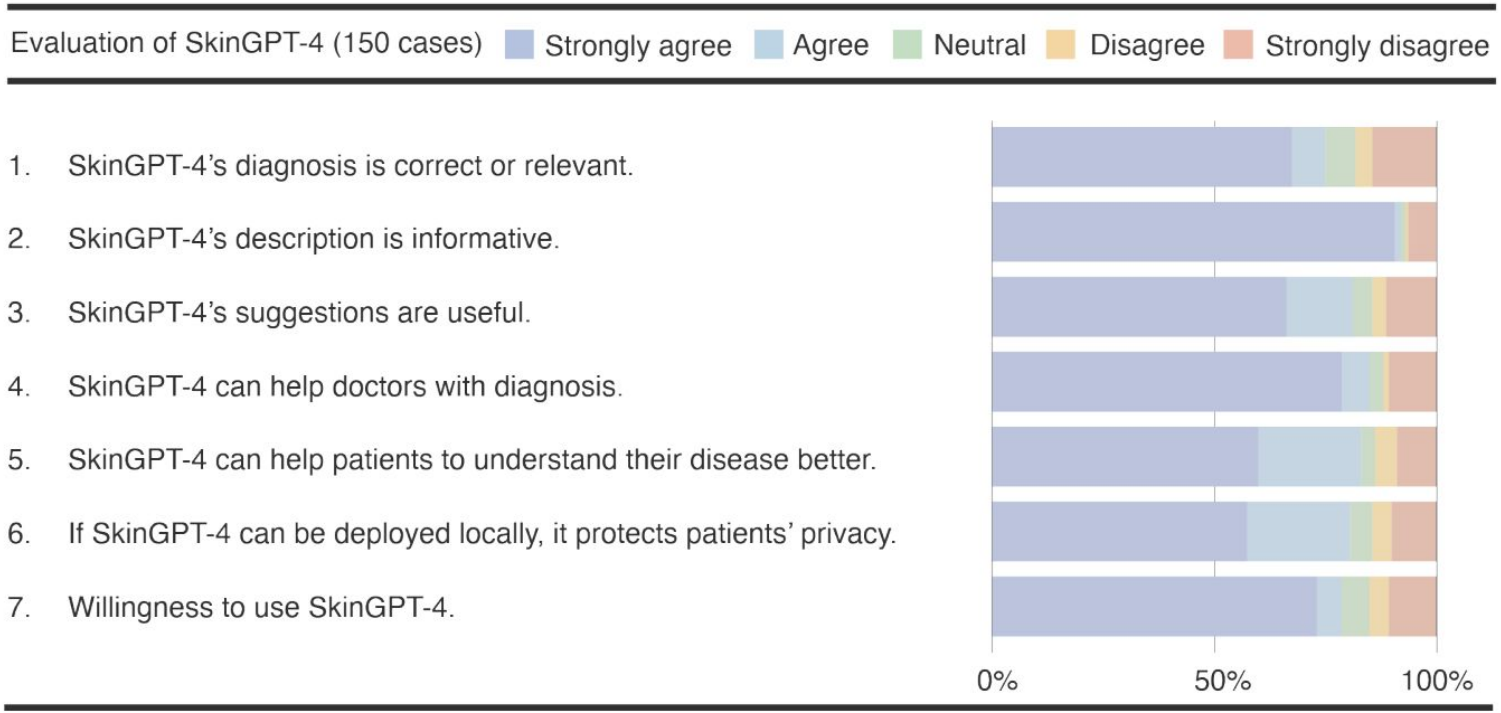
Medical multimodal foundation model based on LLM

- SkinGPT-4



Medical multimodal foundation model based on LLM

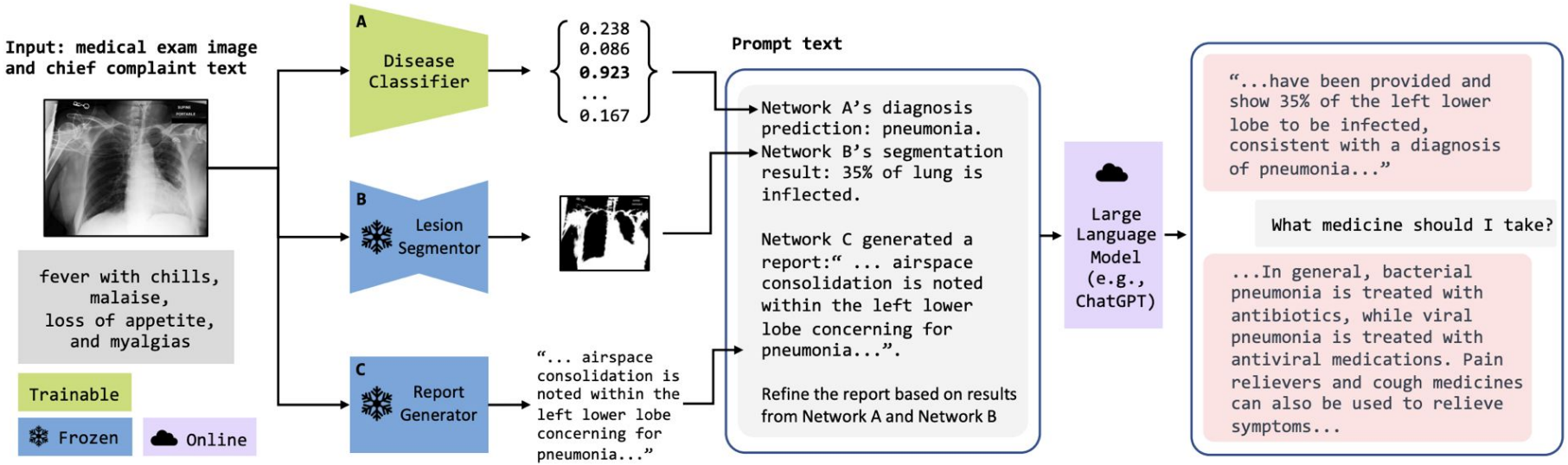
- Clinical evaluation of SkinGPT-4 by certified offline and online dermatologists



Zhou, et al., SkinGPT-4: An Interactive Dermatology Diagnostic System with Visual Large Language Model, arXiv 2023.

Medical multimodal foundation model based on LLM

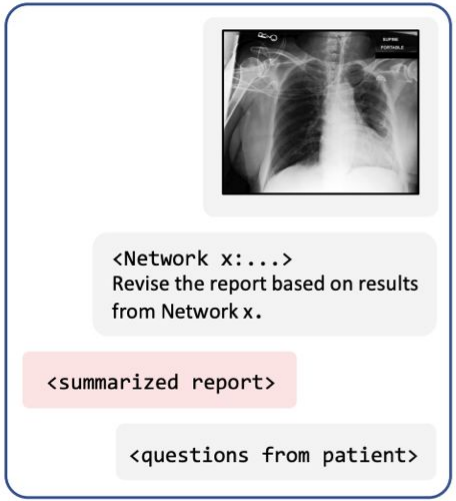
- ChatCAD
  - Use LLMs to enhance the output of multiple CAD networks by summarizing and reorganizing information presented in natural language text format
  - Leverage strengths of LLMs' medical domain knowledge and logical reasoning



Wang, et al., ChatCAD: Interactive computer-aided diagnosis on medical image using large language models, arXiv 2023.

Medical multimodal foundation model based on LLM

- ChatCAD
  - Interactive CAD with LLM

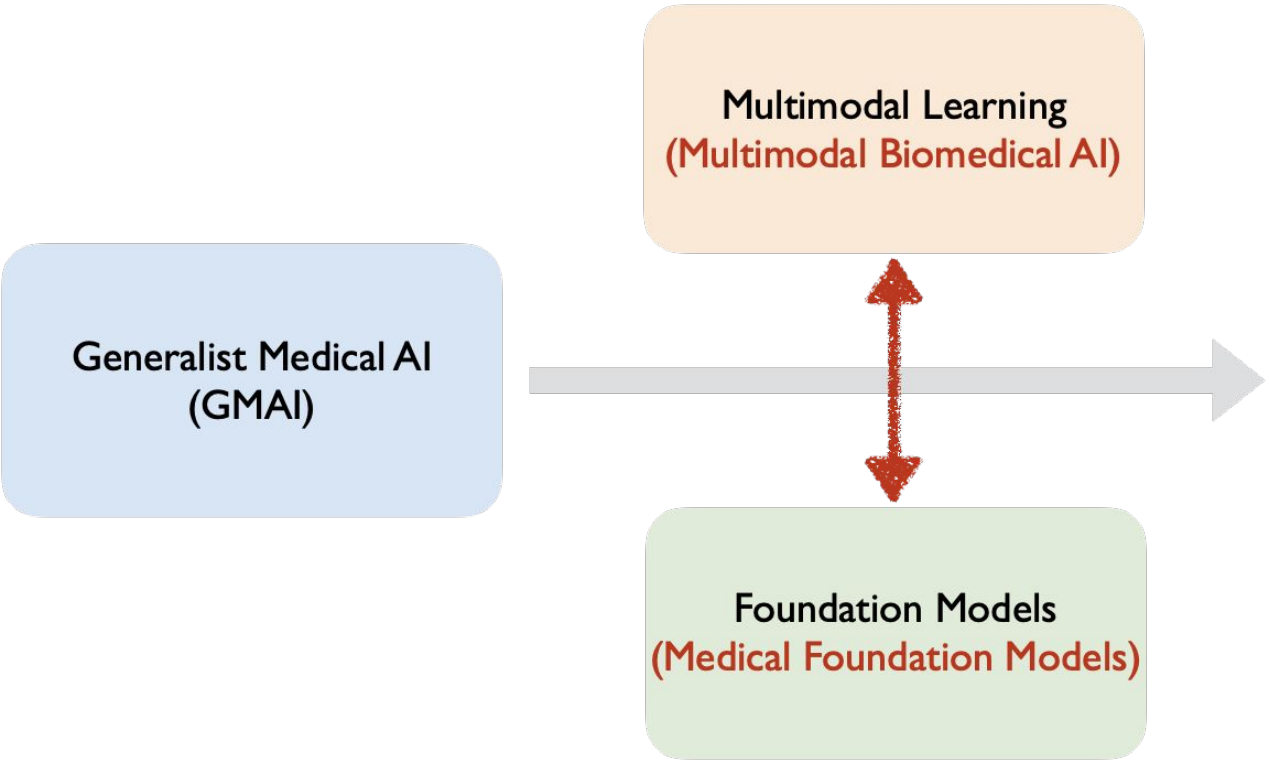


Wang, et al., ChatCAD: Interactive computer-aided diagnosis on medical image using large language models, arXiv 2023.

Today's agenda

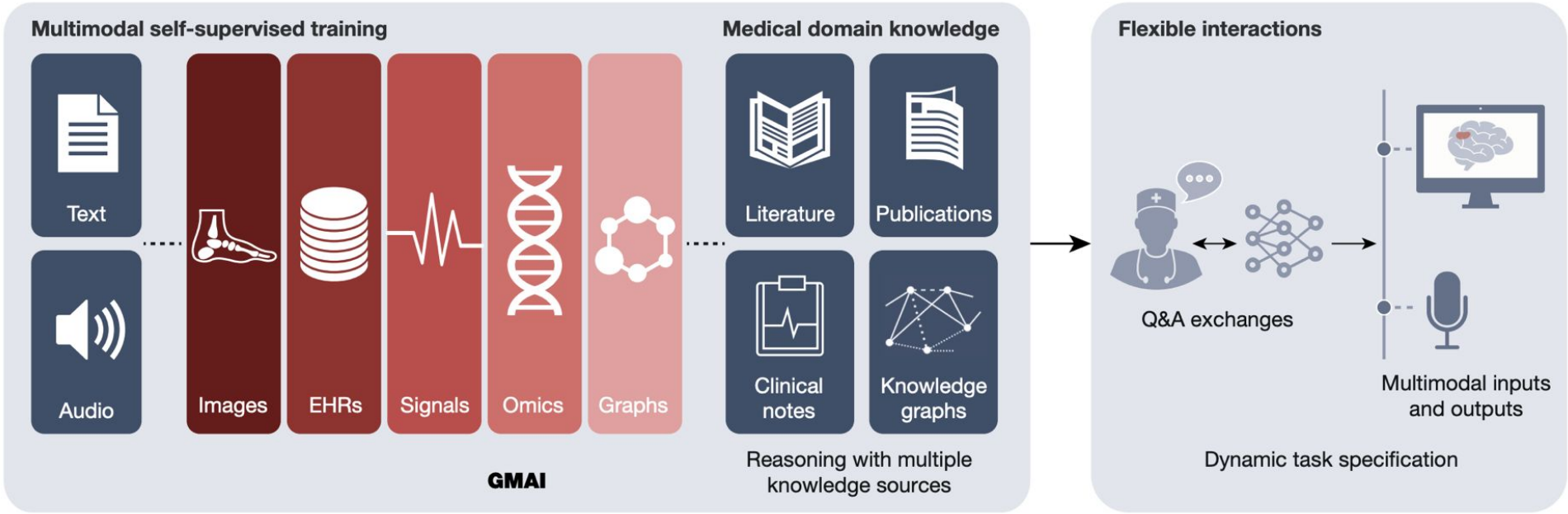
- Foundation model
- Language foundation model
- Vision foundation model
- Multimodal foundation model
- Generalist Medical AI (GMAI)

Generalist Medical AI (GMAI)



Generalist Medical AI (GMAI)

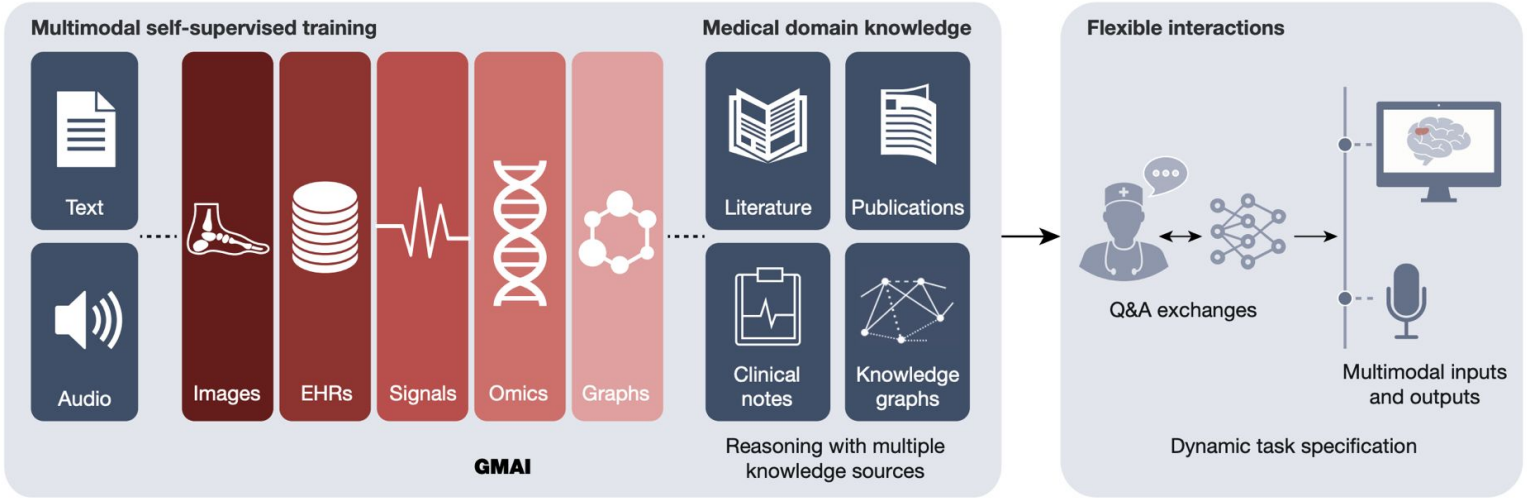
- Generalist paradigm: medical foundation models on multimodal architectures
  - Promise to solve more diverse and challenging tasks than current medical AI models



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

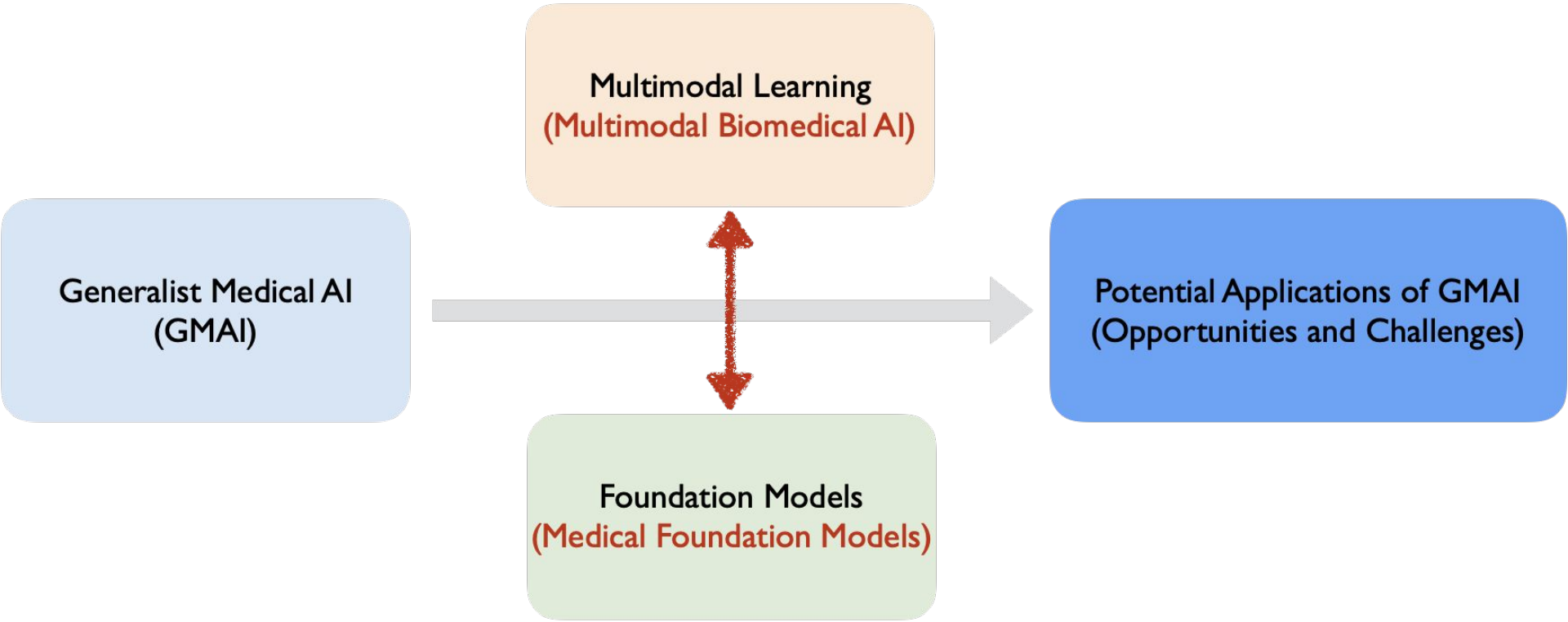
Generalist Medical AI (GMAI)

- Generalist paradigm: medical foundation models on multimodal architectures
  - Adapt to new tasks with dynamic task specification
  - Flexible combinations and interactions of data modalities
  - Represent medical domain knowledge for reasoning new tasks



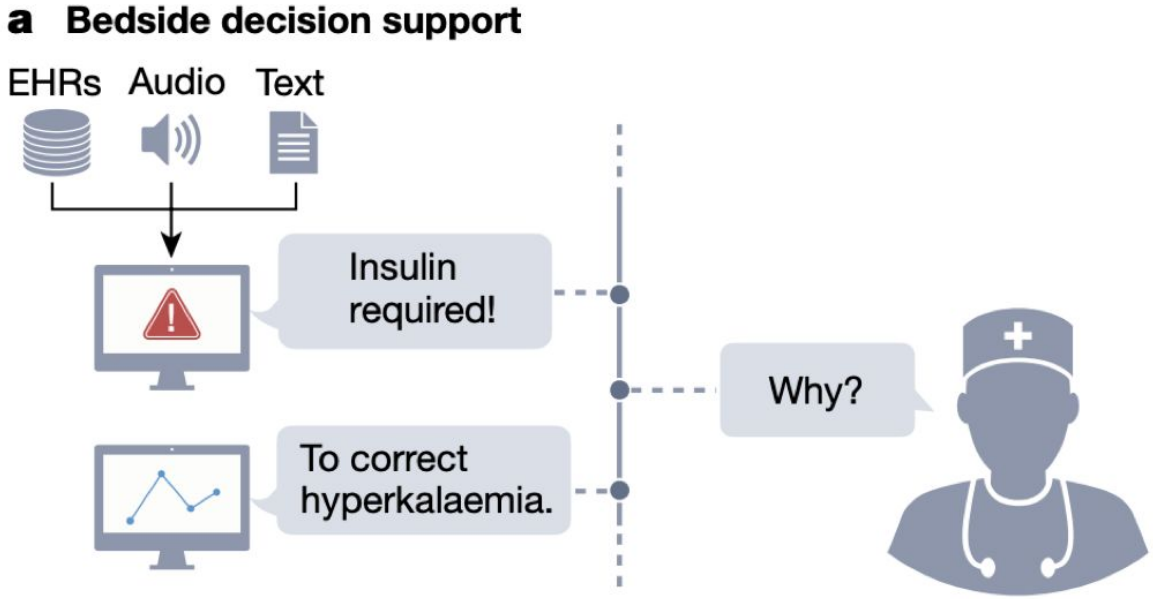
Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

Generalist Medical AI (GMAI)



Potential applications of GMAI

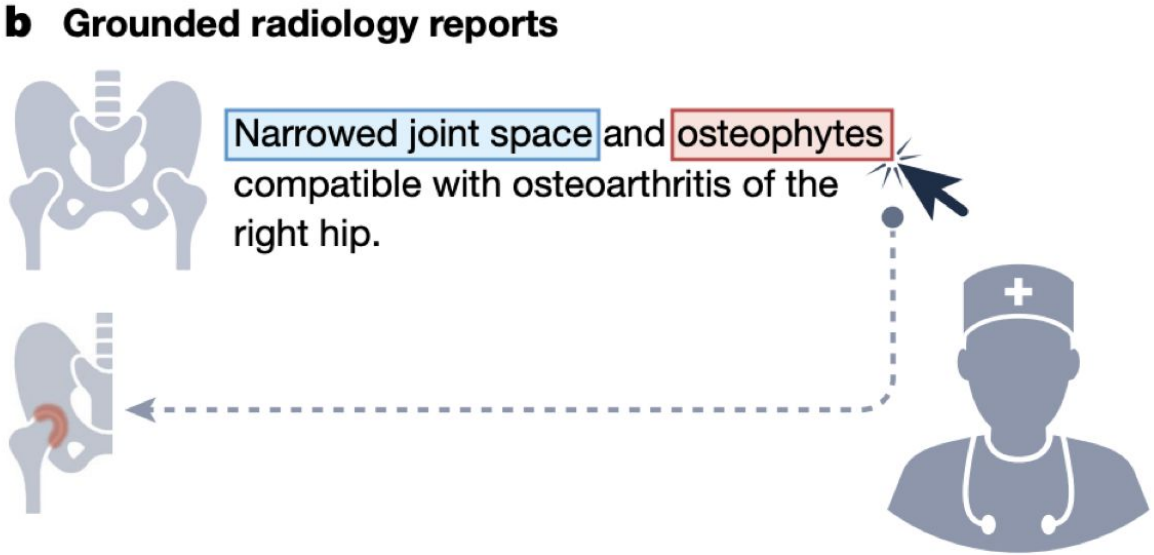
- Generalist paradigm: medical foundation models on multimodal architectures
  - Bedside decision support



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

Potential applications of GMAI

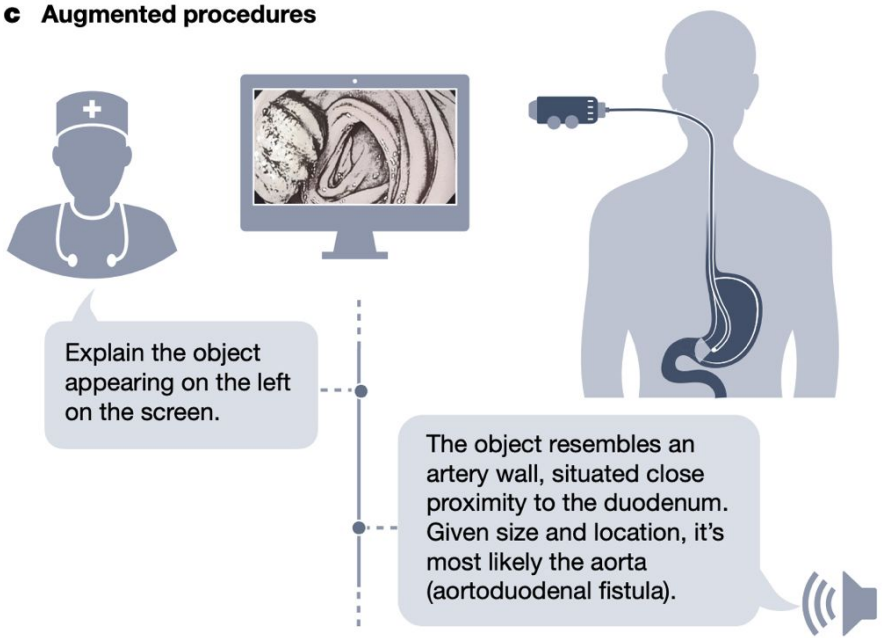
- Generalist paradigm: medical foundation models on multimodal architectures
  - Grounded radiology reports



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

Potential applications of GMAI

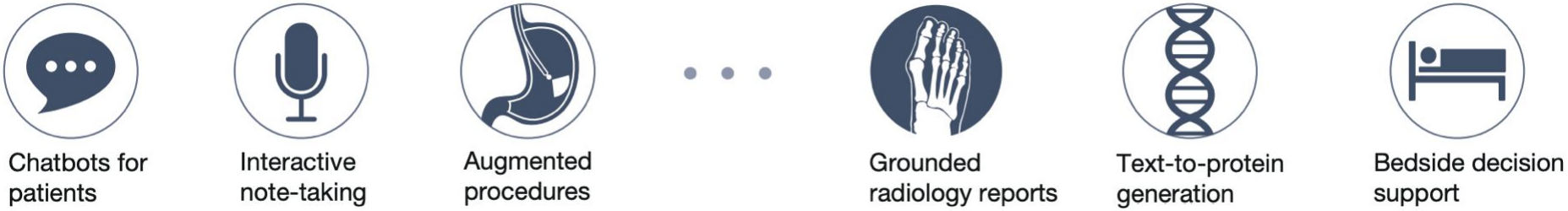
- Generalist paradigm: medical foundation models on multimodal architectures
  - Augmented procedures



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

Potential applications of GMAI

- Generalist paradigm: medical foundation models on multimodal architectures
  - Bedside decision support
  - Grounded radiology reports
  - Augmented procedures
  - Interactive note-taking
  - Chatbots for patients
  - Text-to-protein generation
  - ……



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

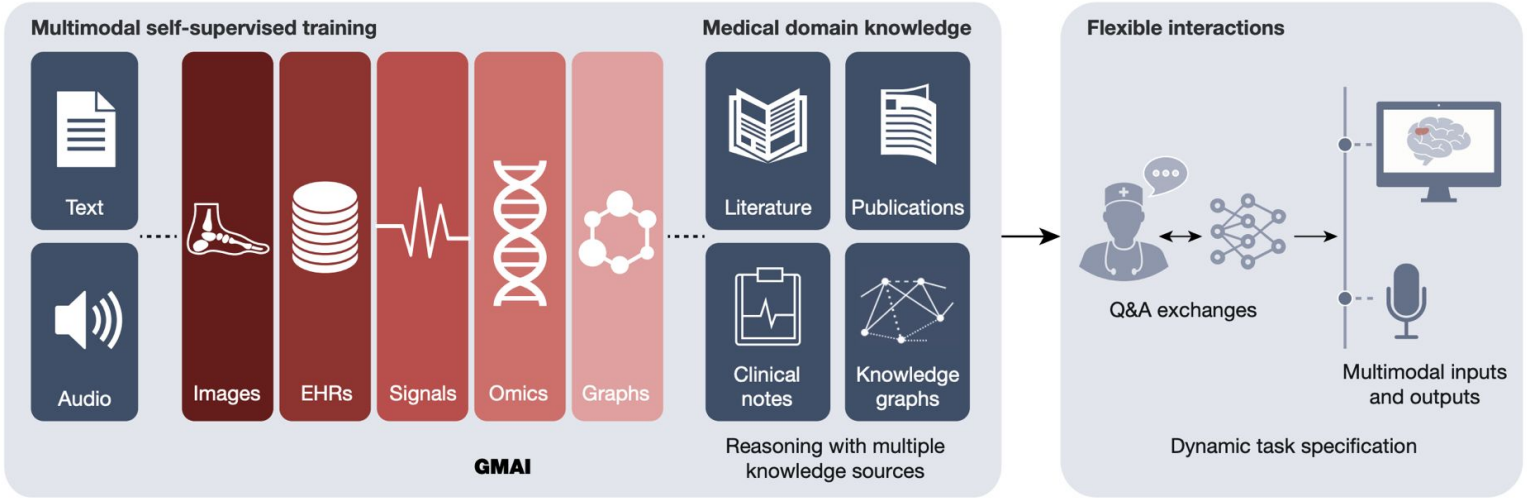
Challenges of GMAI

- Validation
- Verification
- Social biases
- Privacy
- Scale
- ……

Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

Generalist Medical AI (GMAI)

- Generalist paradigm: medical foundation models on multimodal architectures
  - Adapt to new tasks with dynamic task specification
  - Flexible combinations and interactions of data modalities
  - Represent medical domain knowledge for reasoning new tasks



Moor, et al., Foundation models for generalist medical artificial intelligence, Nature 2023.

All lectures are done!

Any feedback and suggestions?

Date	Lecture #	Topic	Papers	Instructor / Presenter
Tue 8/29	1	Introduction and course overview		Liyue Shen
Thu 8/31	2	Biomedical imaging with deep learning **[Fundamental]**		Liyue Shen
Tue 9/5	3	Implicit neural representation learning **[Advanced]**		Liyue Shen
Thu 9/7	4	Generative diffusion models **[Advanced]**		Liyue Shen
Tue 9/12	5	Medical image analysis **[Fundamental]**		Liyue Shen
Thu 9/14	6	Multimodal foundation models **[Advanced]**		Liyue Shen
Mon 9/18		Drop/add deadline for full term classes		
Tue 9/19	7	Implicit neural representation learning		
Thu 9/21	8	Implicit neural representation learning		
Tue 9/26	9	Implicit neural representation learning		
Thu 9/28	10	Implicit neural representation learning		
Tue 10/3	11	Generative diffusion models		
Thu 10/5	12	Generative diffusion models		
Tue 10/10	13	Generative diffusion models		
Thu 10/12	14	Generative diffusion models		
Tue 10/17		No class (fall study break)		
Thu 10/19	15	Self-supervised learning		
Tue 10/24	16	Self-supervised learning		
Thu 10/26	17	Multimodal learning		
Tue 10/31	18	Multimodal learning		
Thu 11/2	19	Transformer and LLM		
Tue 11/7	20	Transformer and LLM		