Current Landscape of Generative AI and Its Applications in Medicine

Kyu-Hwan Jung, Ph.D.

Assistant Professor

Samsung Advanced Institute for Health Science and Technology,
Sungkyunkwan University

Research Institute for Future Medicine,
Samsung Medical Center

01

Current Status of Generative AI

▪ Chatbot as an Interface

GPT-4

You
Please explain the AIRIS 2024 held in Seoul

ChatGPT
The AIRIS 2024 held in Seoul is the "AI Regulatory & International Symposium," a distinguished event that emphasizes the use of Artificial Intelligence (AI) in the development of medical products. This symposium is scheduled to take place from February 26th to 29th, 2024, at THE SHILLA Seoul, in Korea. It is co-organized by the Ministry of Food and Drug Safety (MFDS) of Korea and the U.S. Food and Drug Administration (FDA), showcasing a significant partnership between these two regulatory bodies.

The primary purpose of AIRIS 2024 is to offer a forum for regulatory authorities, industry experts, and academia to engage in in-depth discussions on the challenges and opportunities presented by the integration of AI into medical product development. The symposium aims to foster interactive dialogue that concentrates on technical and scientific issues, emerging regulatory matters, and the sharing of information on how regulators can support the introduction and use of these innovative technologies to ensure the delivery of safe and effective medical products to the market.

The event will focus on various areas within the medical sector, including pharmaceuticals, biologics, and medical devices. It aims to attract participants who are not only eager to learn and exchange information about the role of AI within the regulatory landscape but are also actively dealing with challenges related to AI in their regulatory reviews and decision-making processes. The symposium aspires to be a catalyst for advancing the understanding and application of AI in the development of medical products, ensuring that regulatory practices evolve to keep pace with technological advancements [""] [""].

Gemini

AIRIS 2024 행사에 대해서 알려줘

답안 표시

AIRIS 2024 행사 정보

주제: 의료기기 개발에 인공지능 활용

일시: 2024년 2월 26일 (월) ~ 29일 (목)

장소: 서울 THE SHILLA

주최: 대한민국 식품의약품안전처, 미국 식품의약국

참가 대상: 규제 기관, 산업 및 학계 전문가

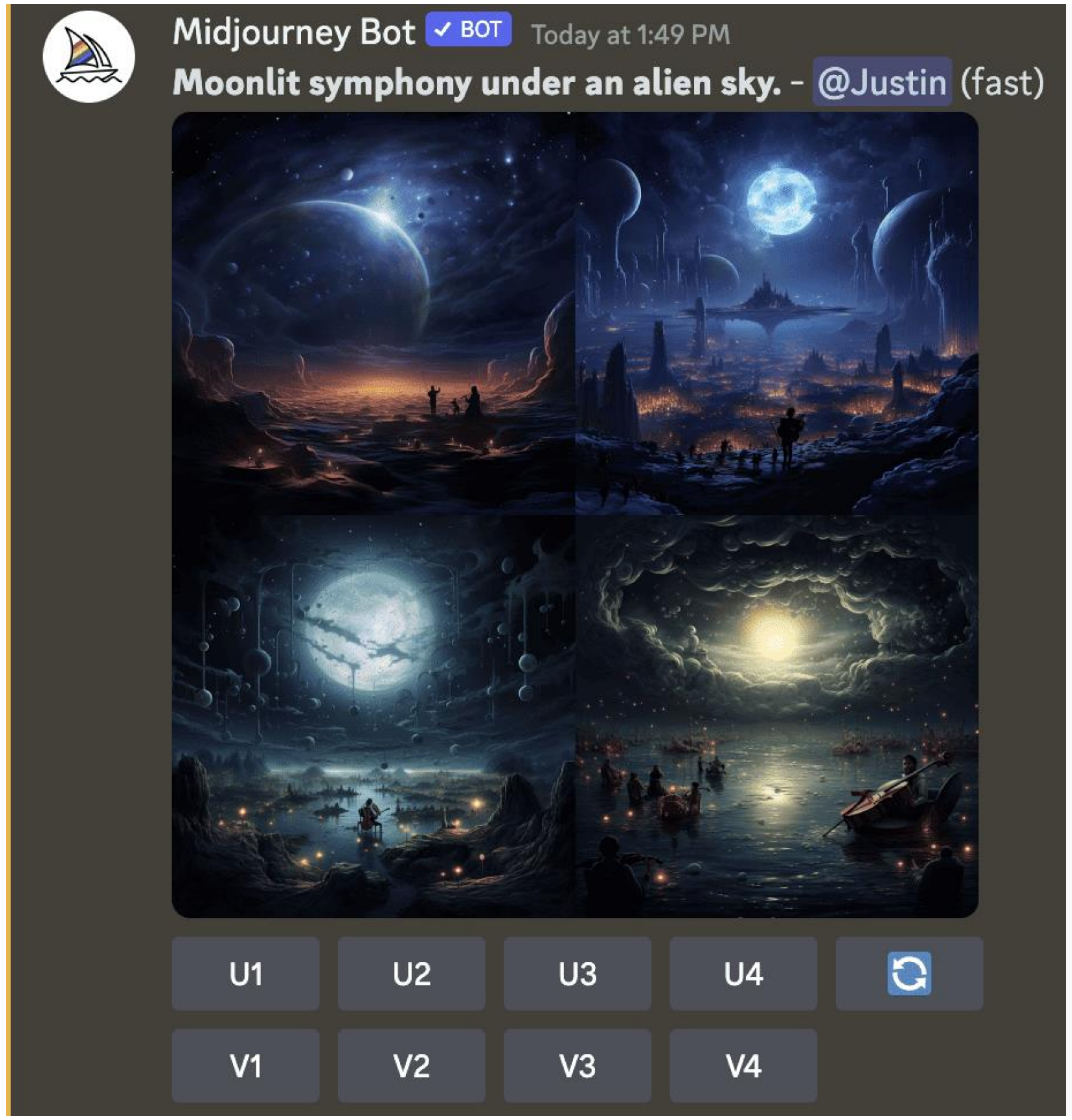
행사 프로그램:

- **2월 26일 (월):** 현장 방문, 개막식
- **2월 27일 (화):**
  - 기조연설: 인공지능 기술 활용 현황 및 전망
  - 세션 1: 의료기기 개발에서 인공지능 기술 도입
  - 만찬 (연사 및 규제기관 대상)
- **2월 28일 (수):**
  - 세션 4: 인공지능 활용 의료기기 분야의 규제 경험
  - 오찬 (연사 및 규제기관 대상)
  - 세션 5: 인공지능 기술 활용한 새로운 의료기기 개발과 치료
  - 만찬 (연사 및 규제기관 대상)
- **2월 29일 (목):**
  - 세션 6: 인공지능 기술 활용의 윤리 및 사회적 책임
  - 폐막식

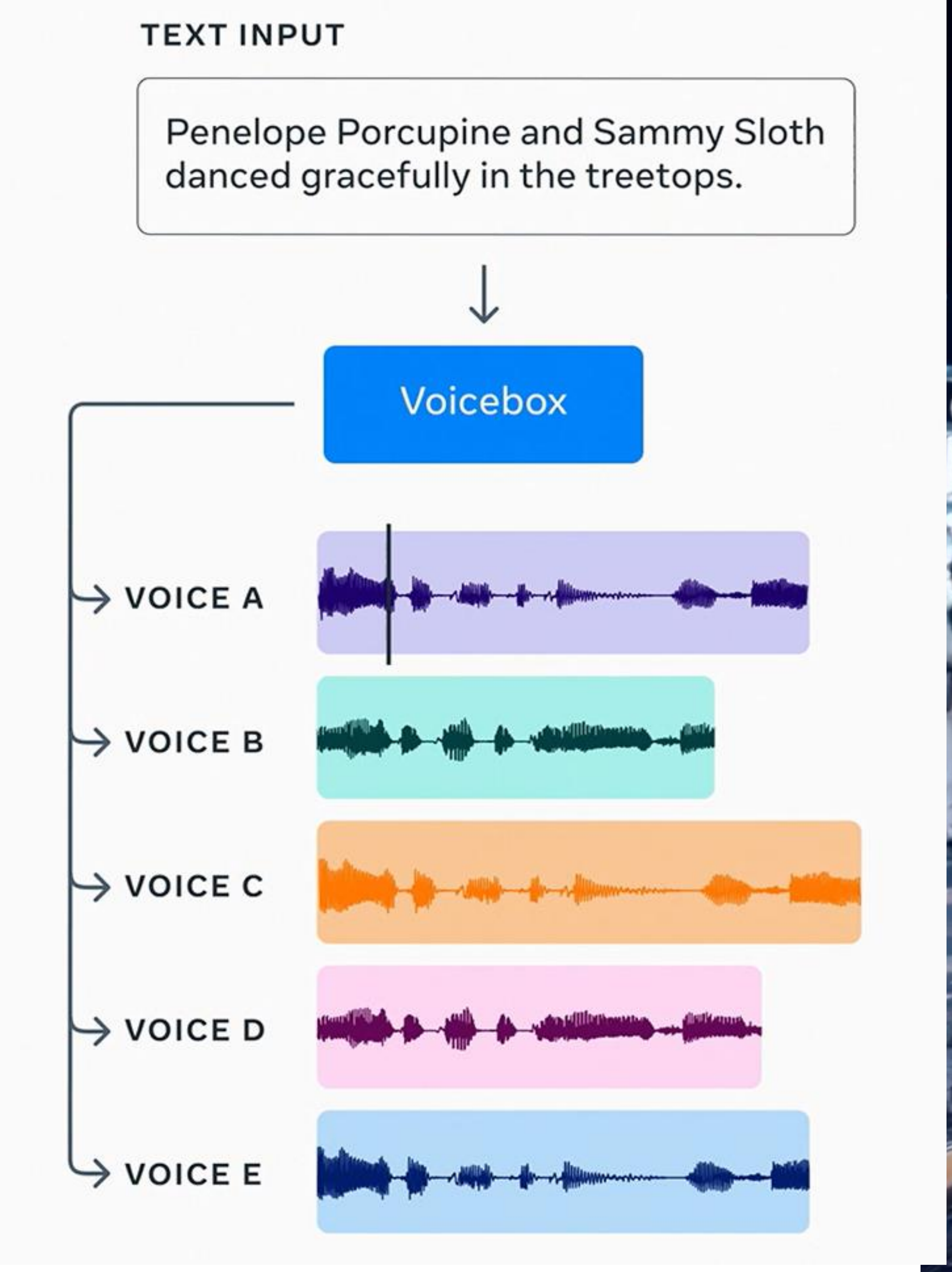
OpenAI(2024), Google(2024)

Generative AI Market Growth

Text to Image



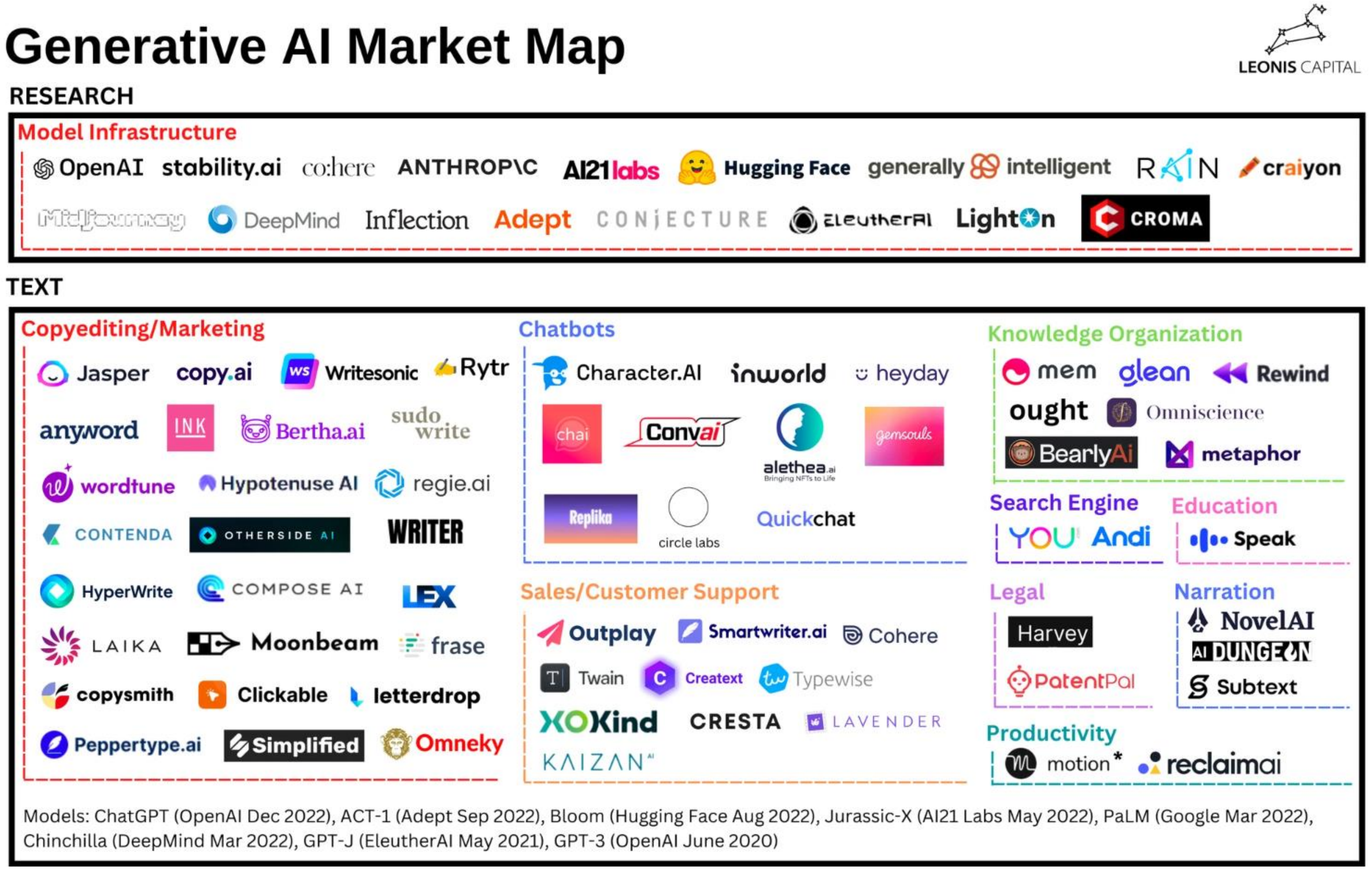
Text to Audio

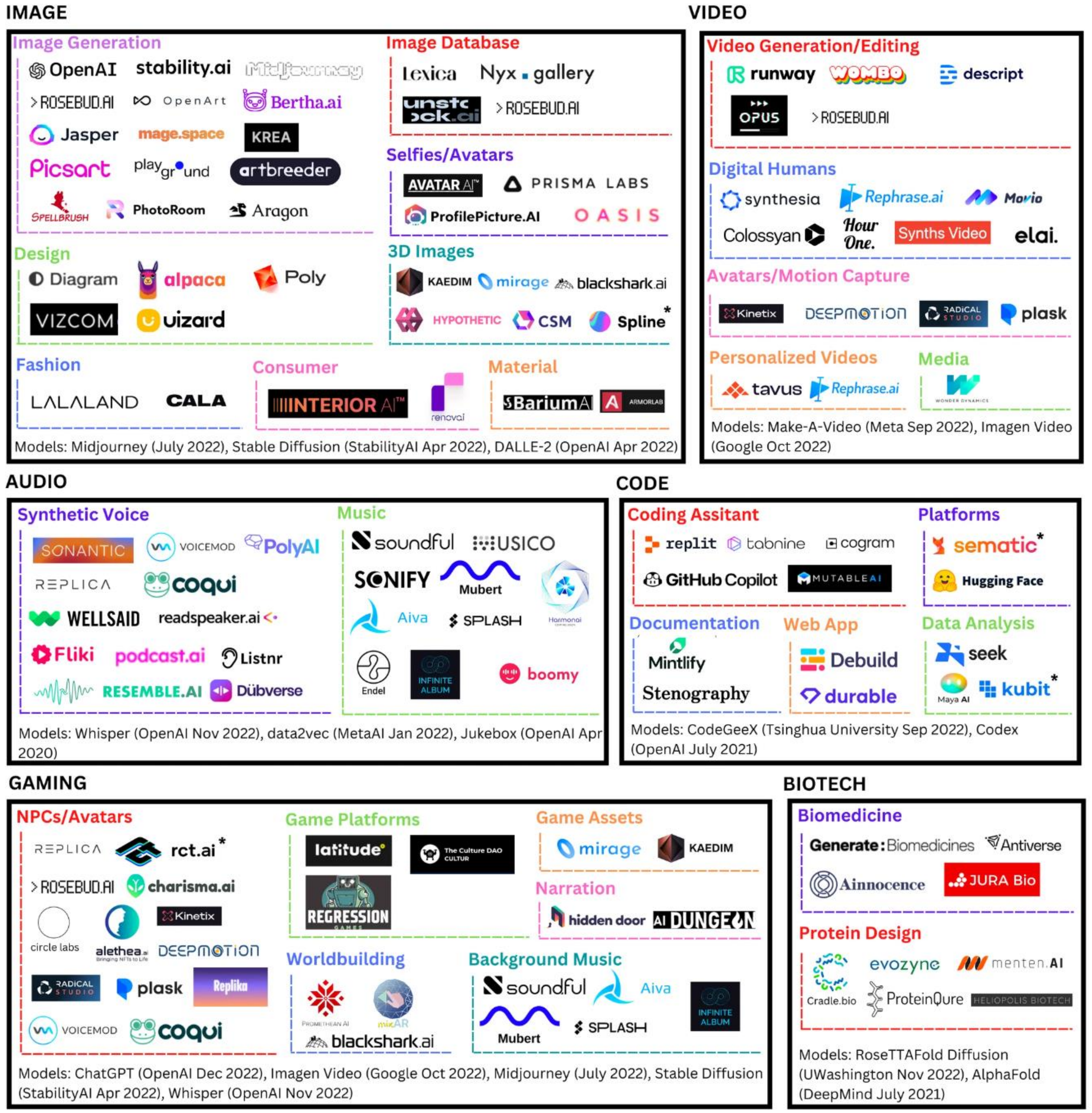


Text to Video

Market.us(2023), towards healthcare(2023)

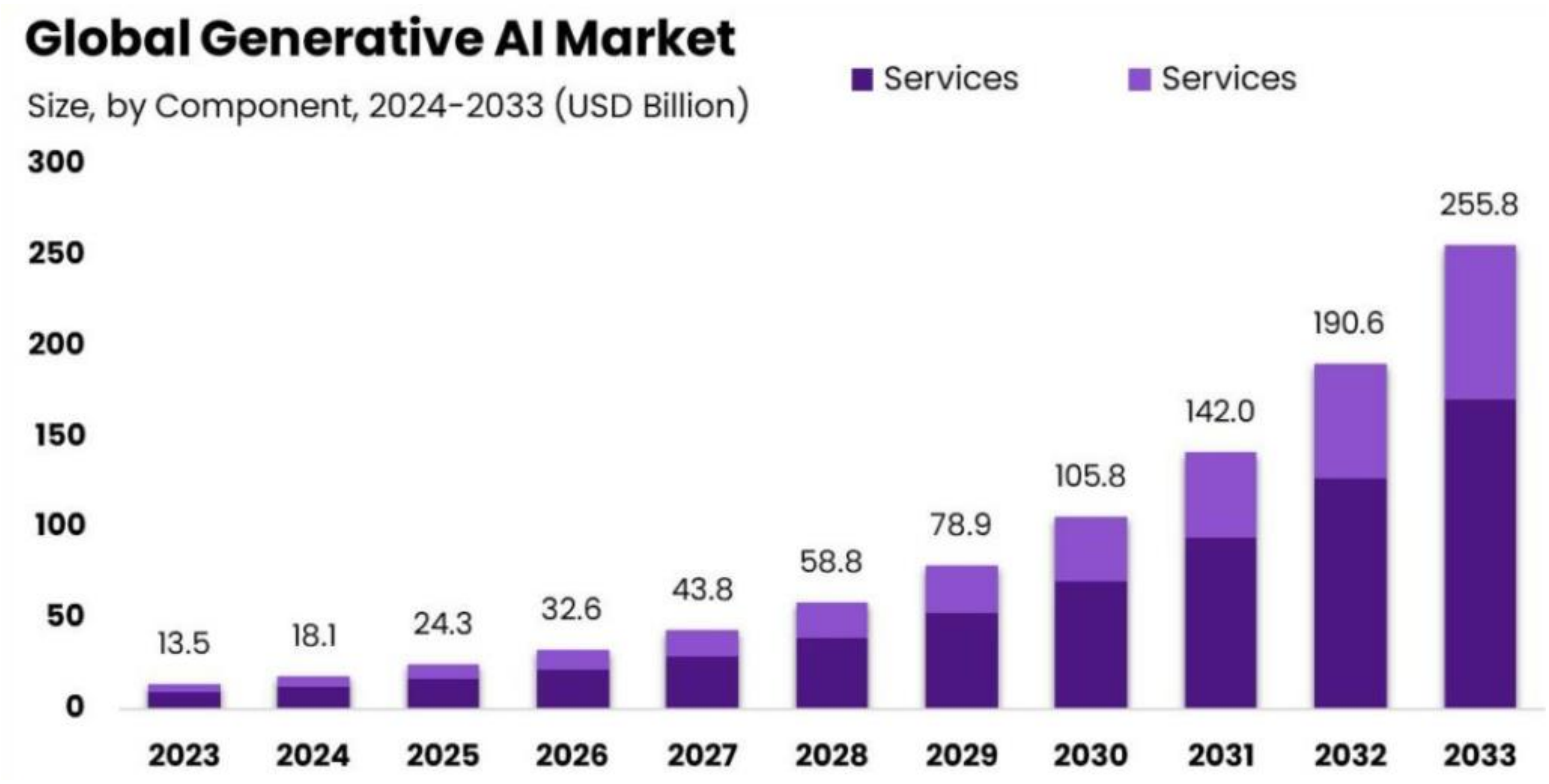
▪ The Landscape of Generative AI Services

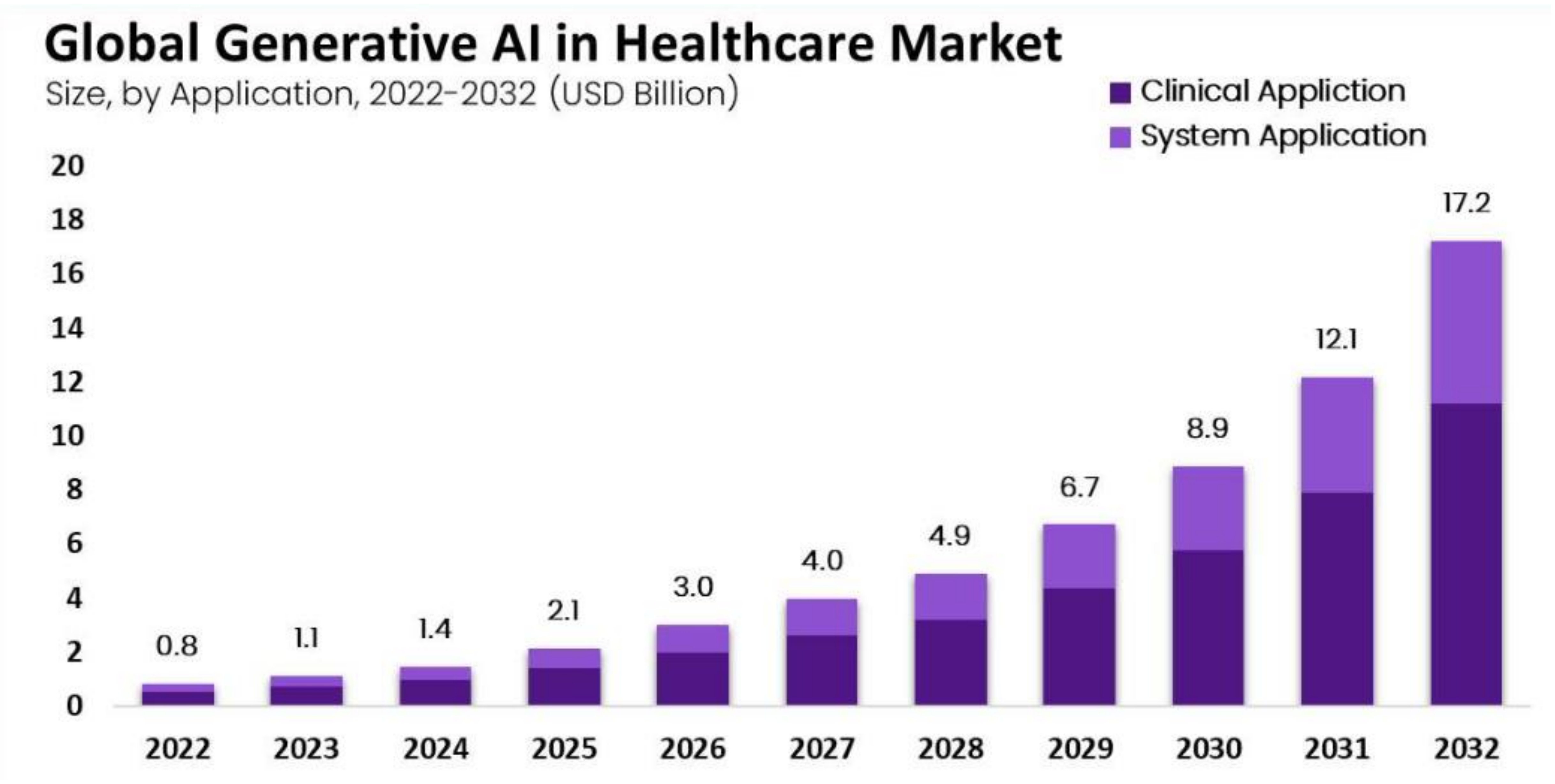




Leonis Capital(2022)

▪ Generative AI Market Growth



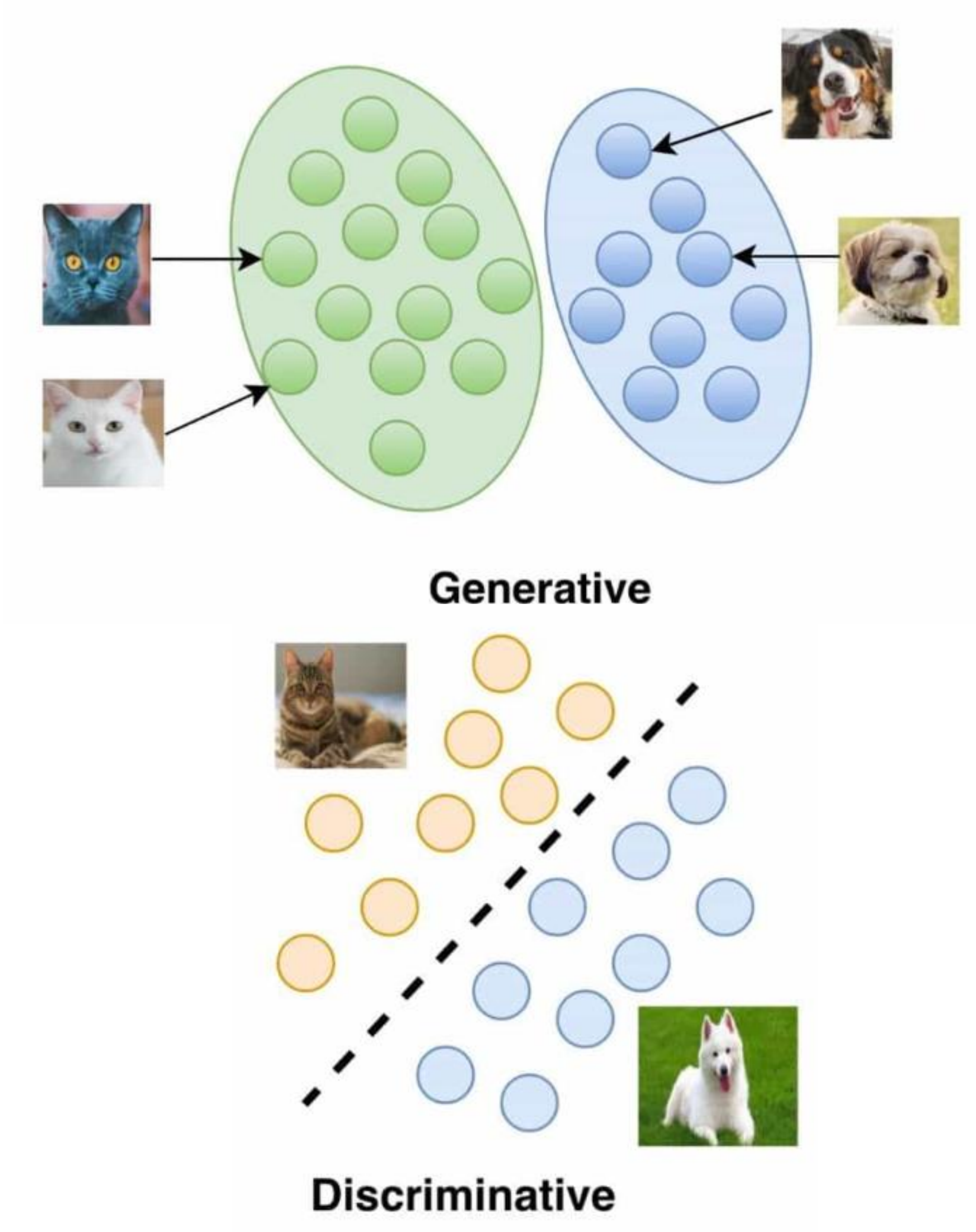


Market.us(2023)

▪ Generative AI vs Discriminative AI

- Generative artificial intelligence is artificial intelligence capable of generating text, images or other data using generative models, often in response to prompts.
- Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.

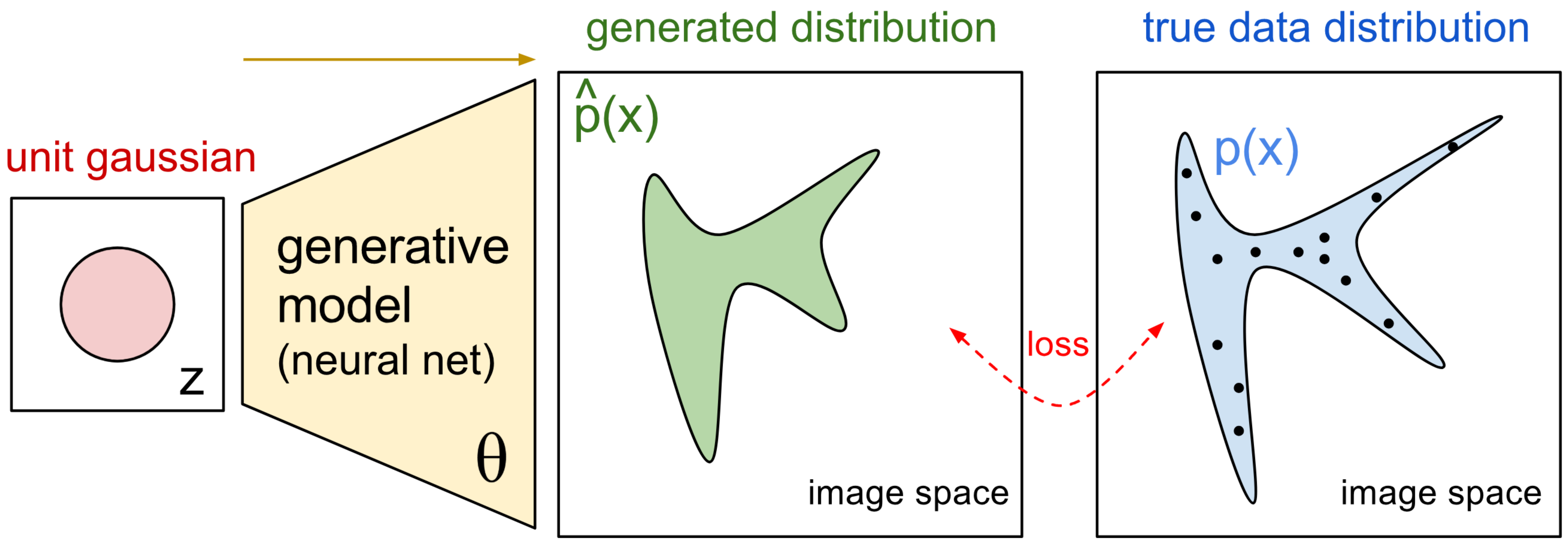
	Generative Models	Discriminative Models
Objective	Model data distribution to generate new samples	Learn decision boundary for accurate classification
Training Approach	Unsupervised /Self-supervised Learning	Supervised /Semi-supervised Learning
Type of Learning	Probabilistic Modeling	Discriminative Modeling
Data Generation	Can generate new samples resembling training data	No inherent data generation capabilities
Decision Boundary	Capture complex decision boundary indirectly	Learn explicit decision boundary between different classes
ML Models	Markov chains, Naïve Bayes, GMM	Logistic regression, SVM, CRF, Decision trees



https://learnopencv.com/generative-and-discriminative-models/

Training Generative AI using Neural Network

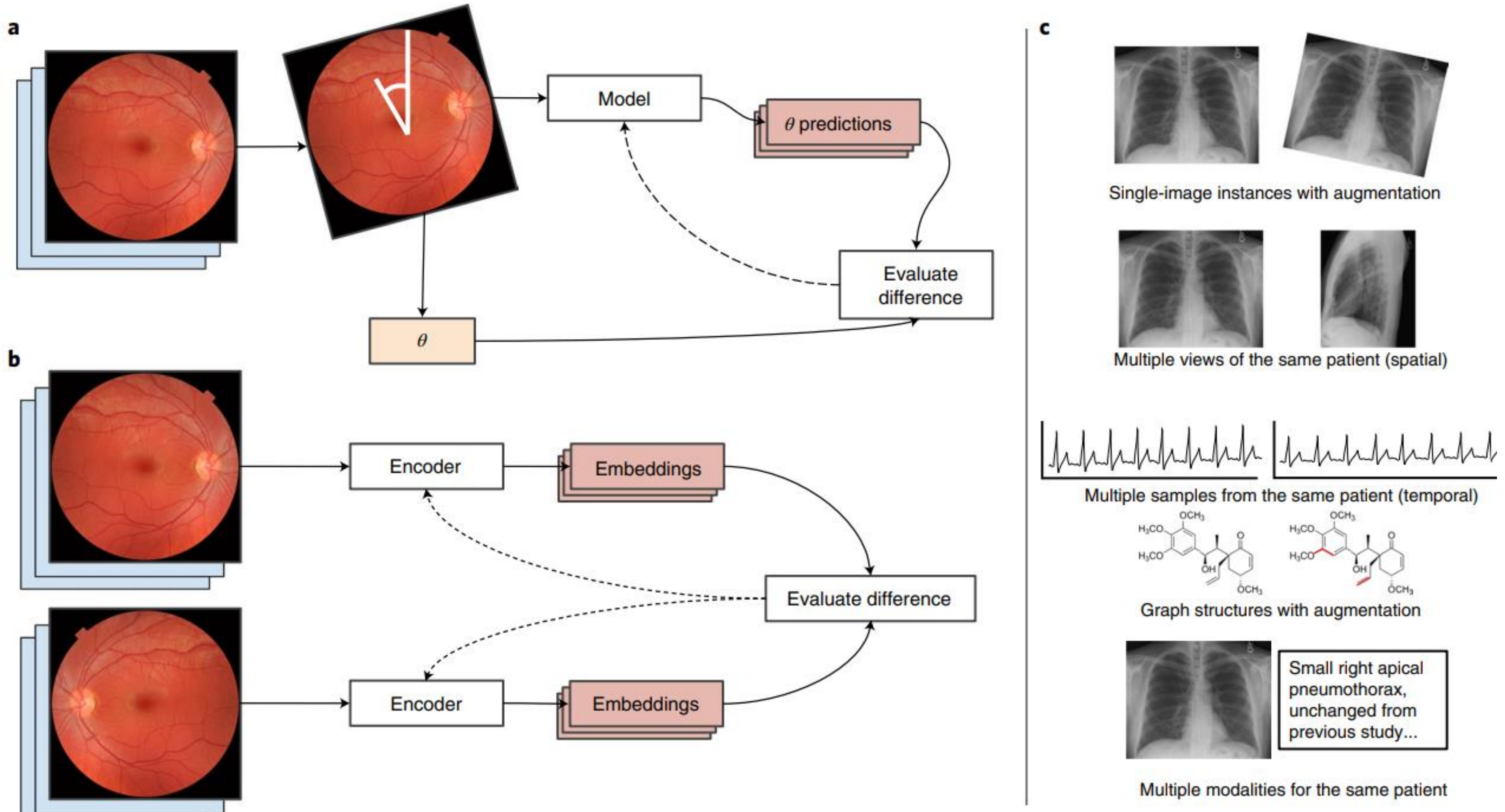
- Given dataset examples $x_1, \ldots, x_n$ from true data distribution $p(x)$, we train a neural network with parameter θ, which map data points from known distribution (such as Gaussian) to some predicted distribution $\hat{p}_\theta(x)$ by minimizing some loss (i.e. KL divergence) between $\hat{p}_\theta(x)$ and $p(x)$.



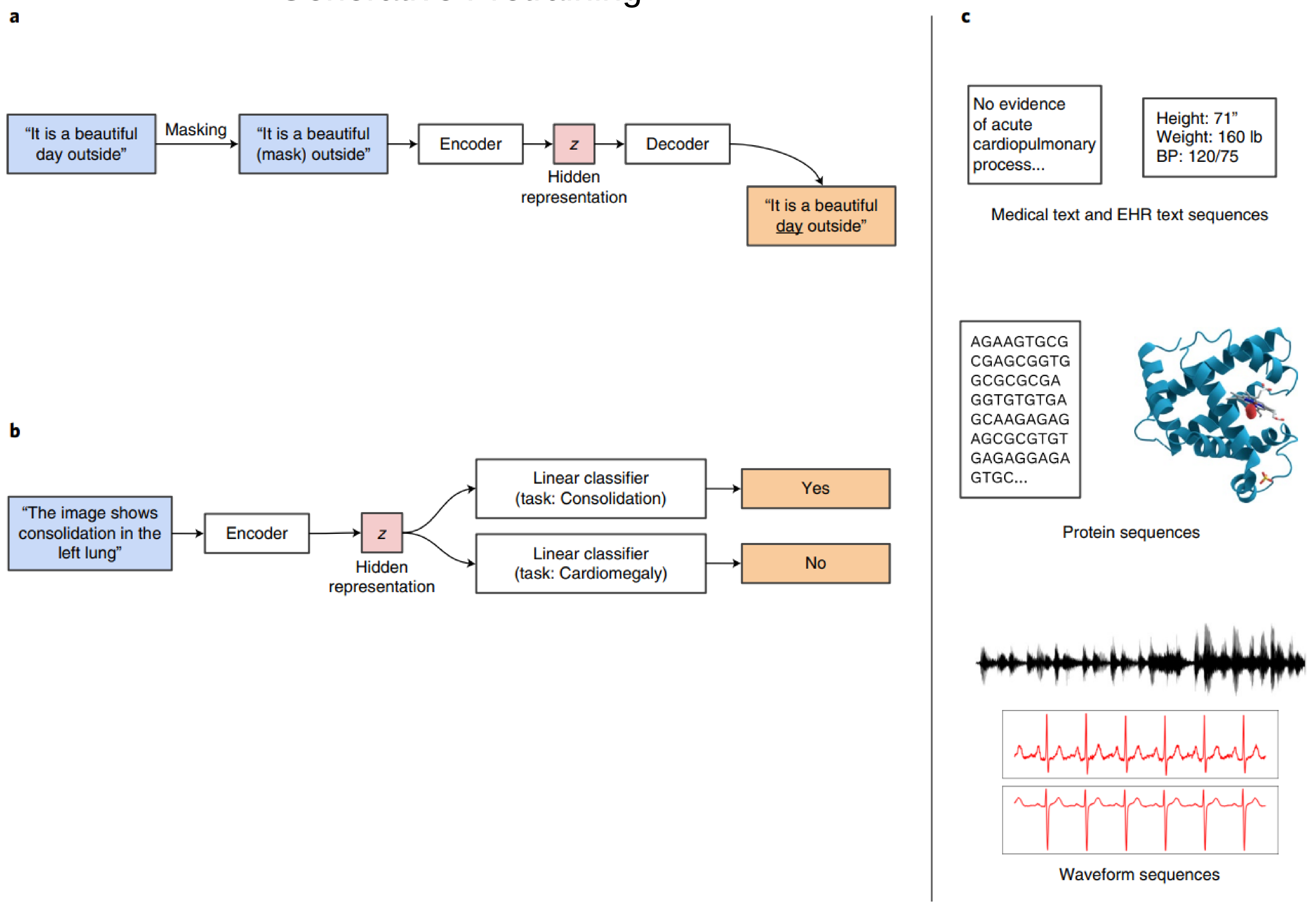
https://blog.openai.com/generative-models/

▪ Training Generative AI using Self-supervised Learning

Contrastive Learning

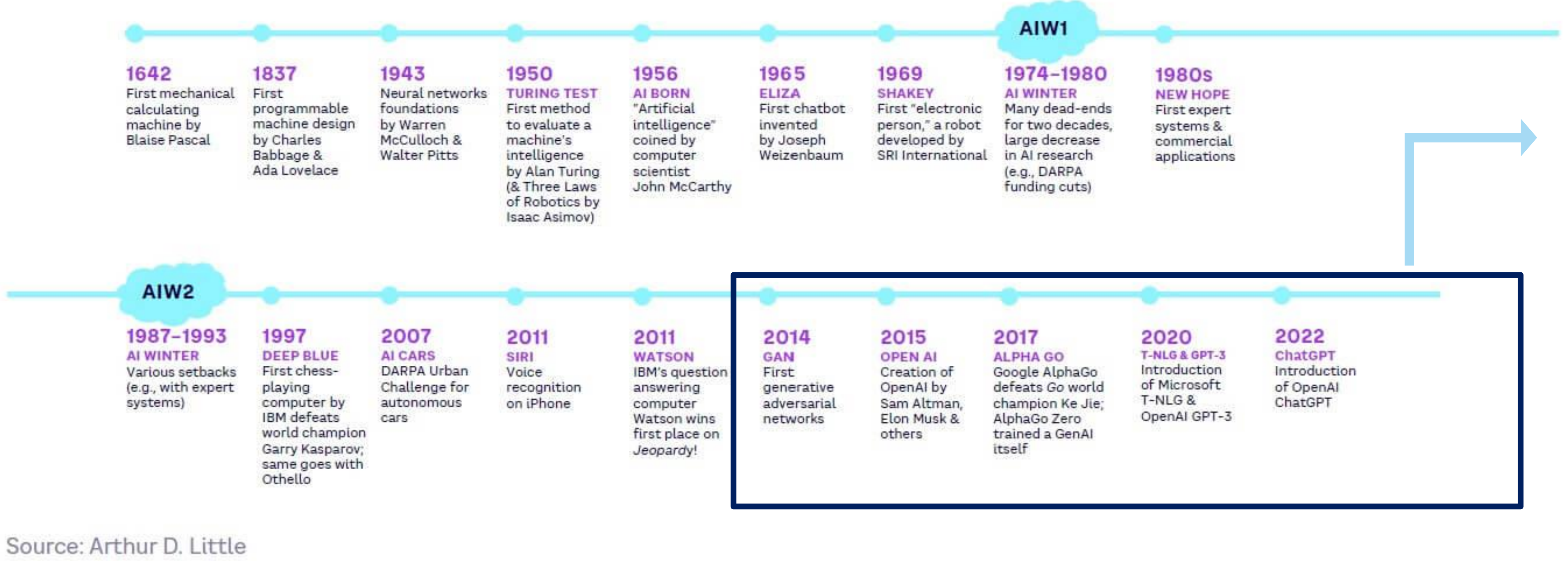


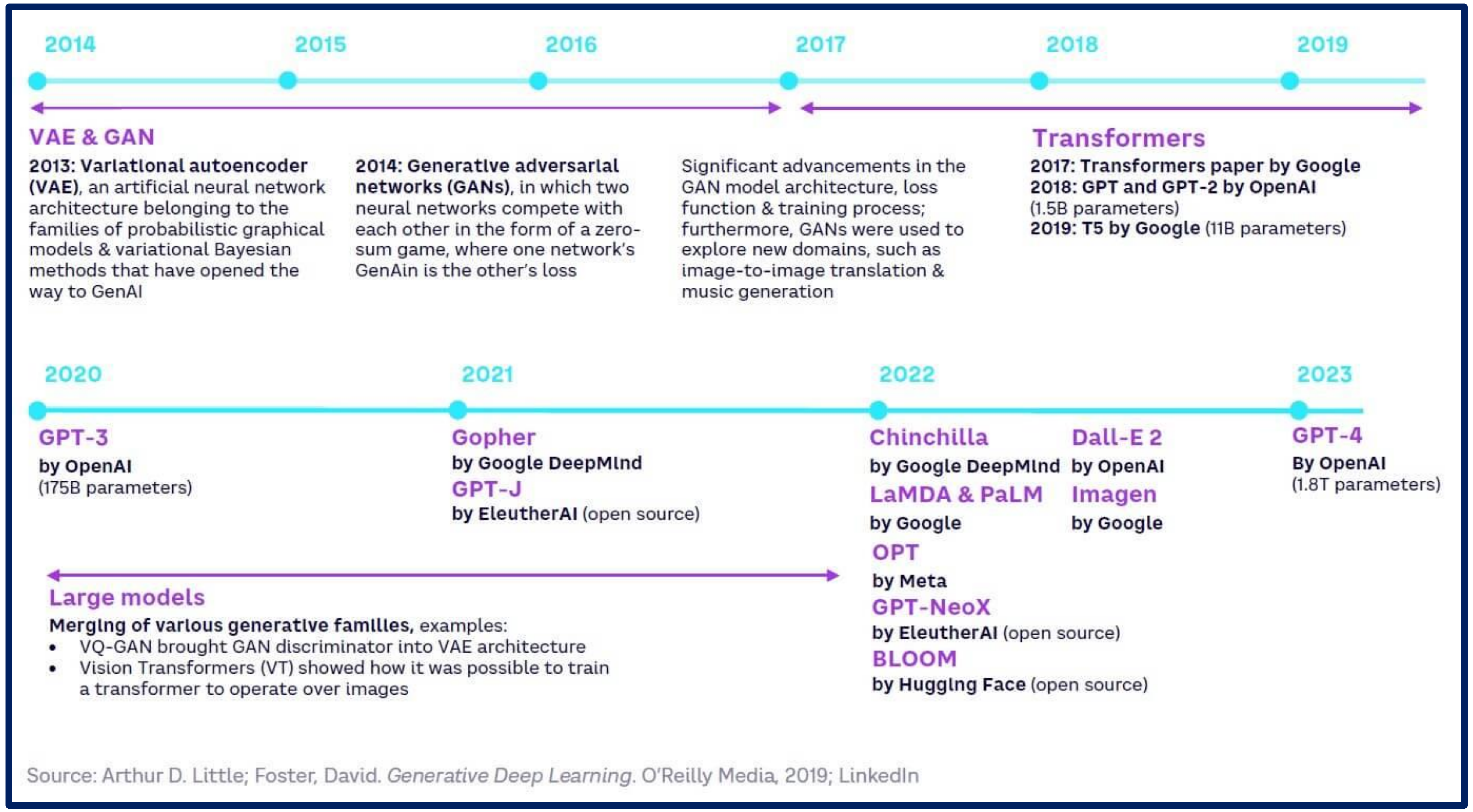
Generative Pretraining

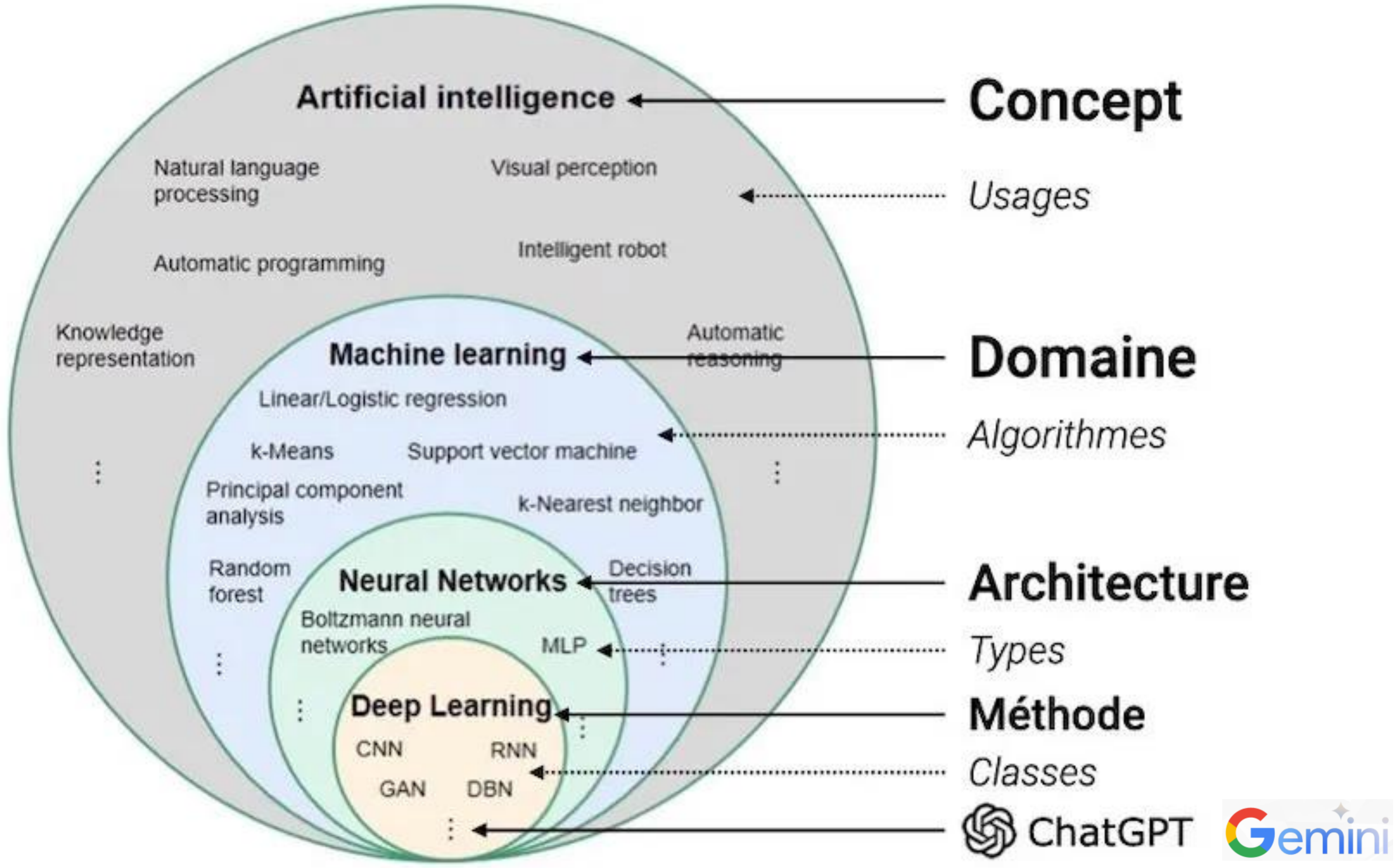


R. Krishnan et al., *Nature Biomedical Engineering*(2022)

▪ History of AI and Generative AI





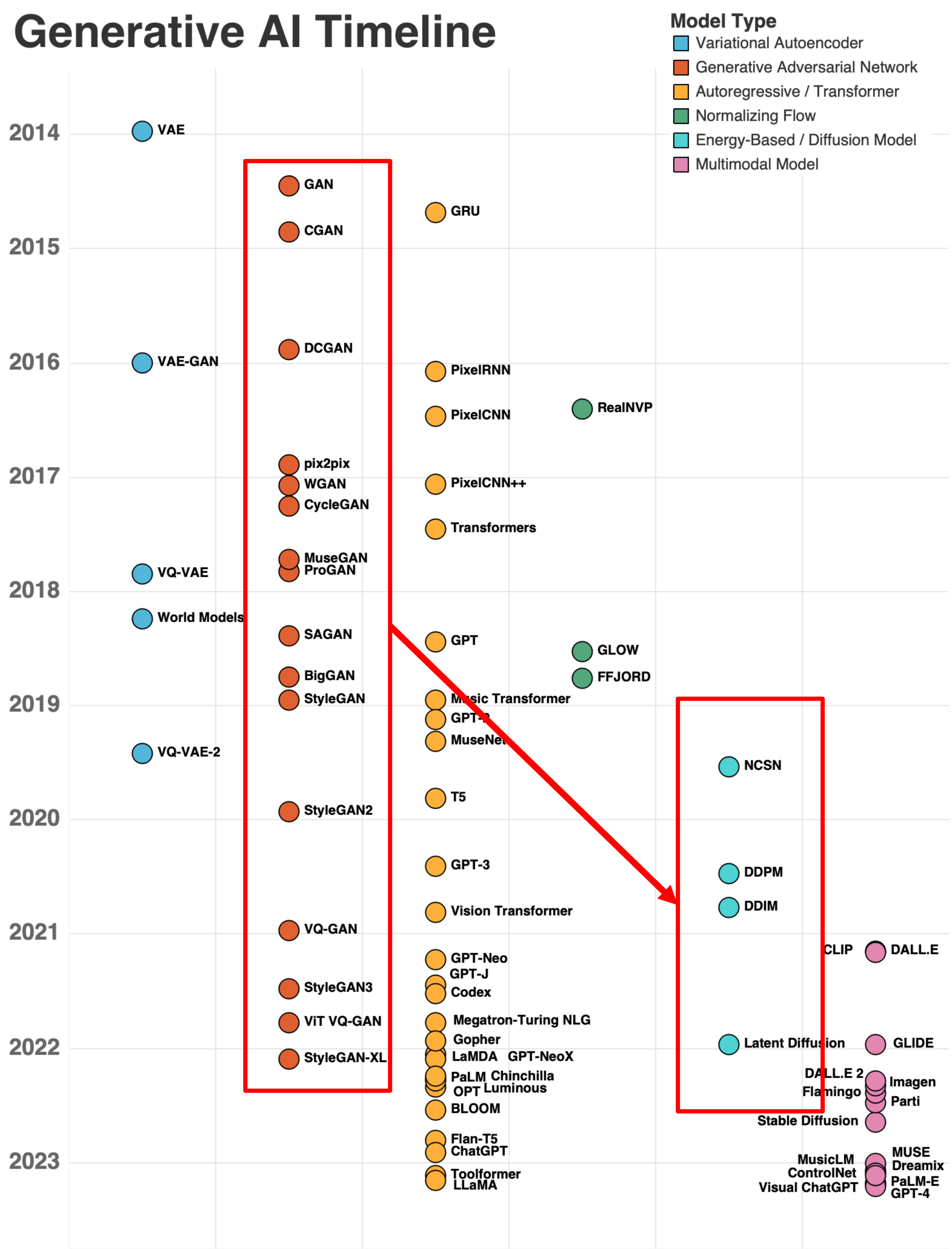


https://www.adlittle.com/cn-en/insights/report/generative-artificial-intelligence-toward-new-civilization

Deep Learning Algorithms for Generative AI

- **Generative Adversarial Networks(GAN)**
  - A framework of estimating generative models via adversarial process of simultaneously training a generative model capturing data distribution and discriminative model to distinguish real and generated data.

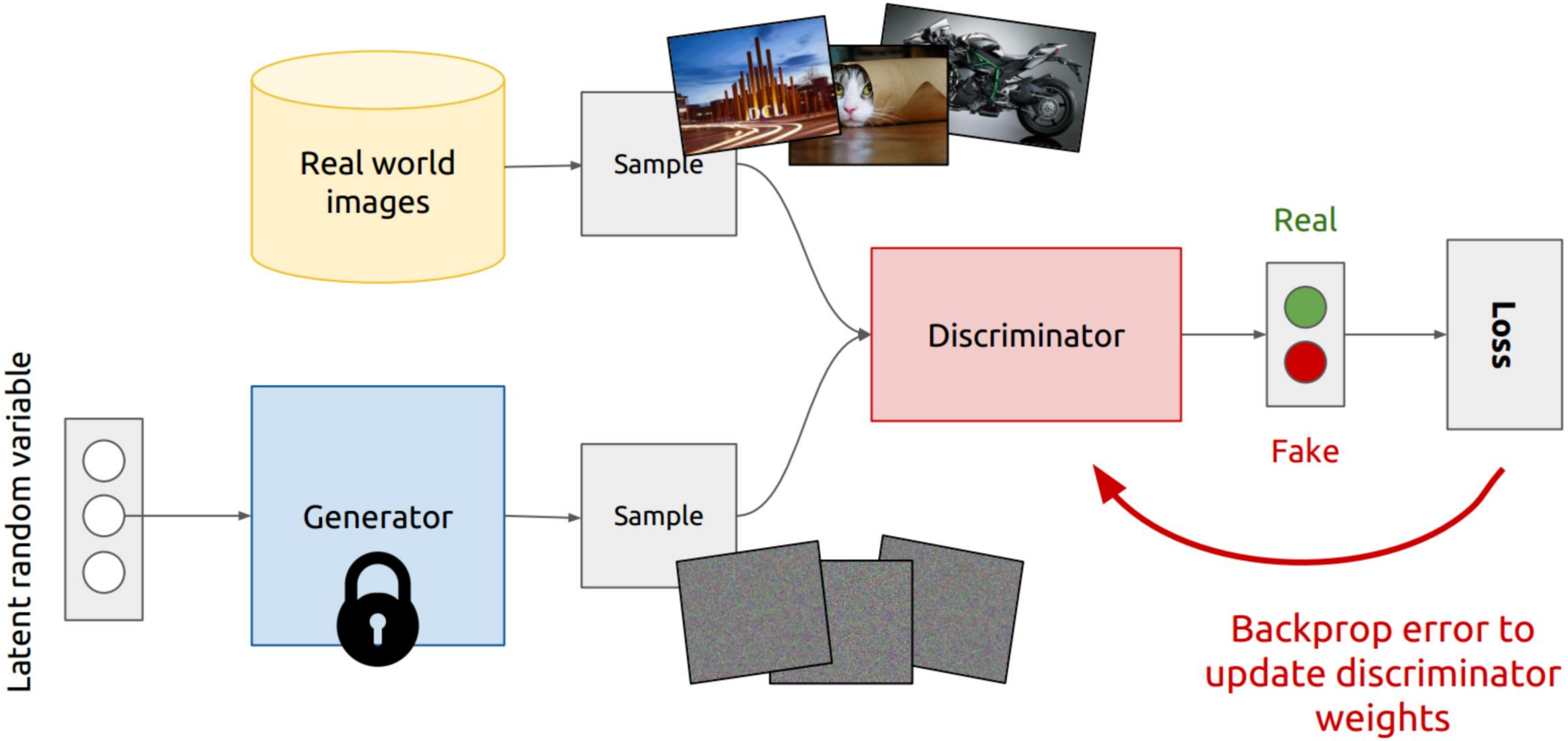
- **Denoising Diffusion Probabilistic Models(DDPM)**
  - A diffusion probabilistic models which is a class of latent variable models inspired by considerations from nonequilibrium thermodynamics.



https://github.com/davidADSP/Generative_Deep_Learning_2nd_Edition

Generative Adversarial Network

- Fix the generator, then train discriminator to distinguish samples of real images from samples synthesized by the generator

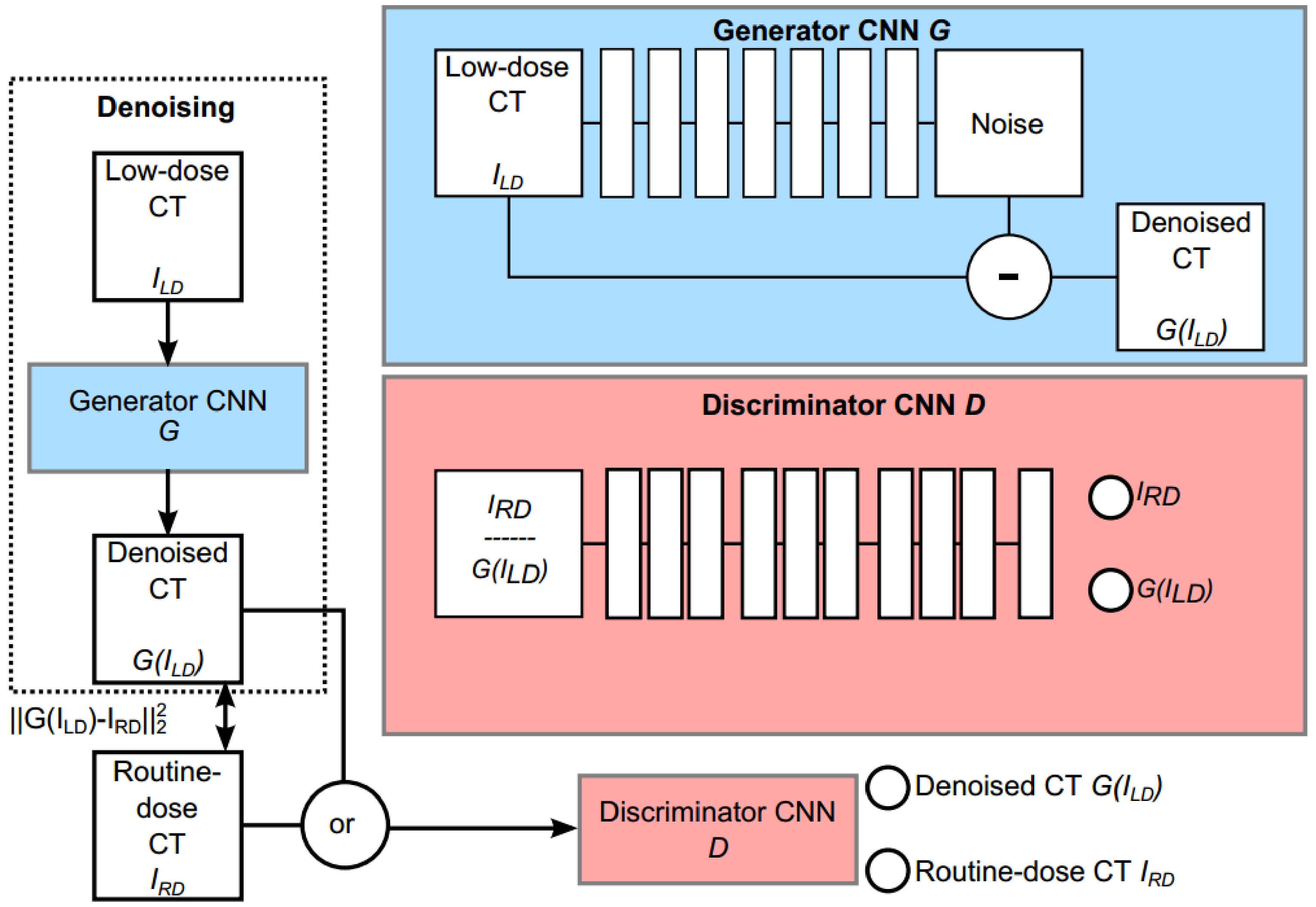


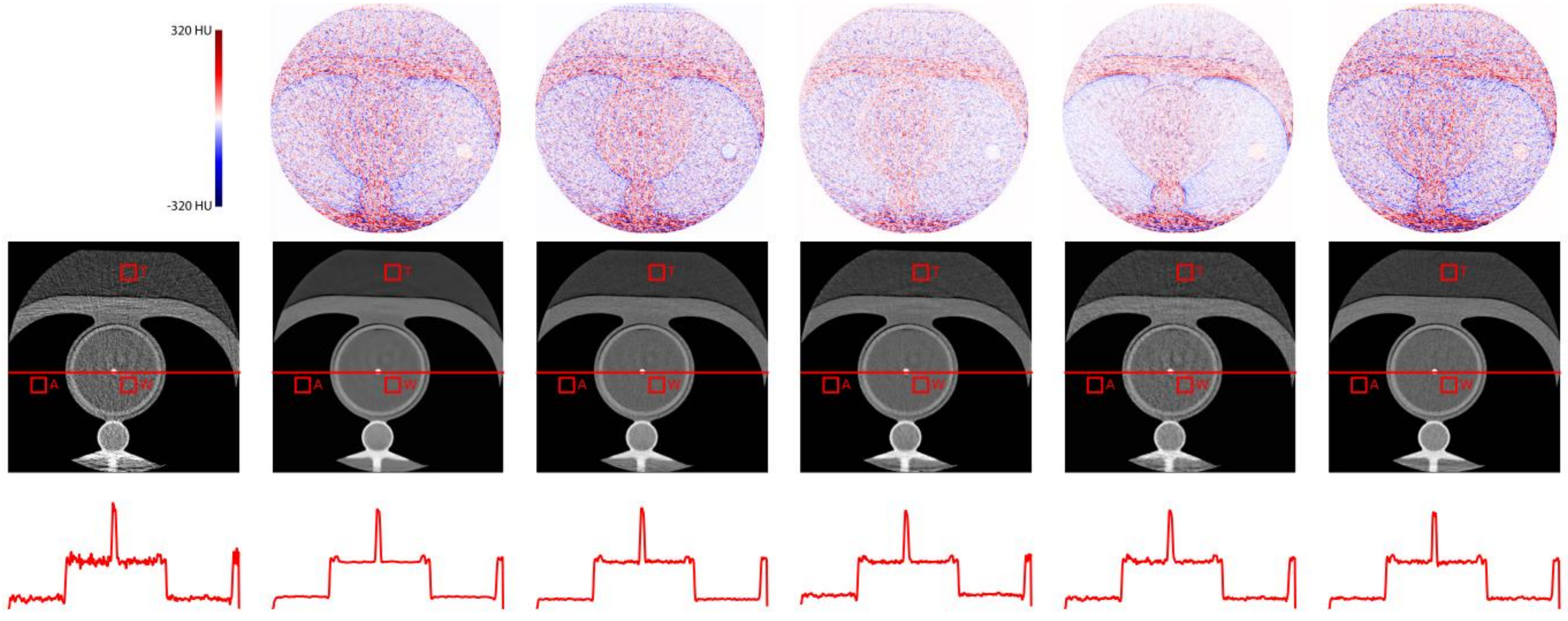
https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

▪ Enhance the quality of low-dose CT to normal-dose CT

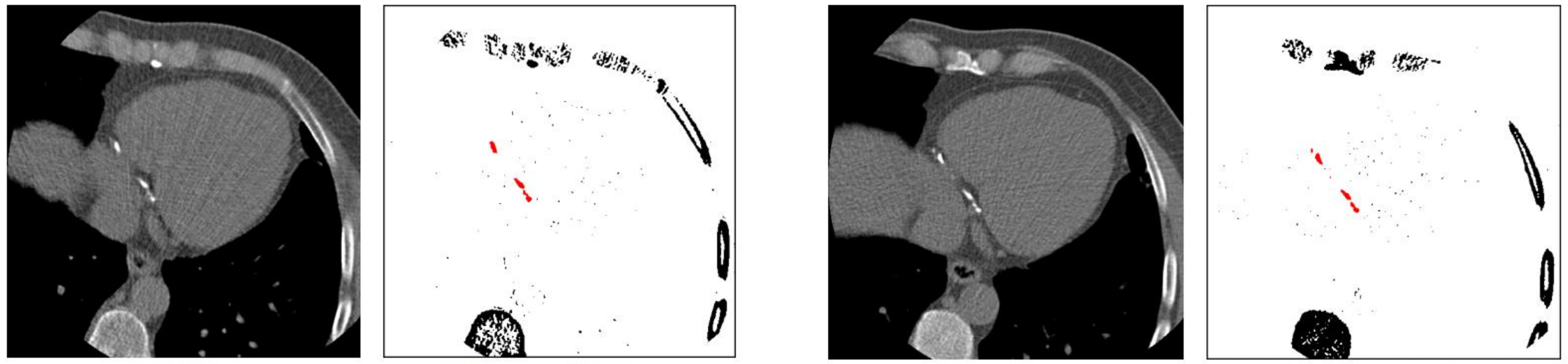
$$\ell_D = \ell_{bce}(D(I_{RD}), 1) + \ell_{bce}(D(G(I_{LD})), 0),$$

$$\ell_G = \lambda_1 ||G(I_{LD}) - I_{RD}||_2^2 + \lambda_2 \ell_{bce}(D(G(I_{LD})), 1)$$





(a) I_{LD} (b) $G_1(I_{LD})$ (c) $G_2(I_{LD})$ (d) $G_3(I_{LD})$ (e) $IR(I_{LD})$ (f) I_{RD}

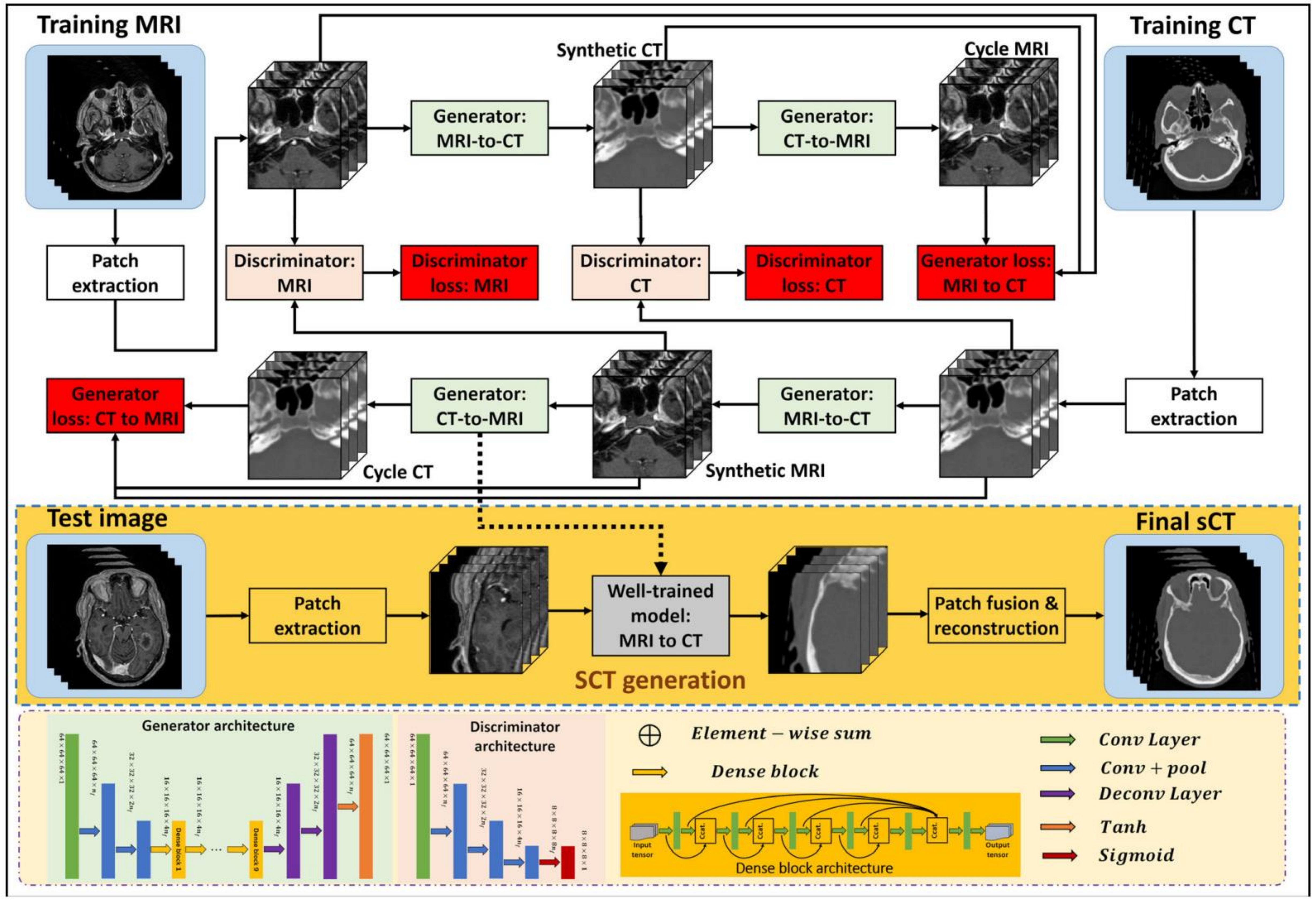


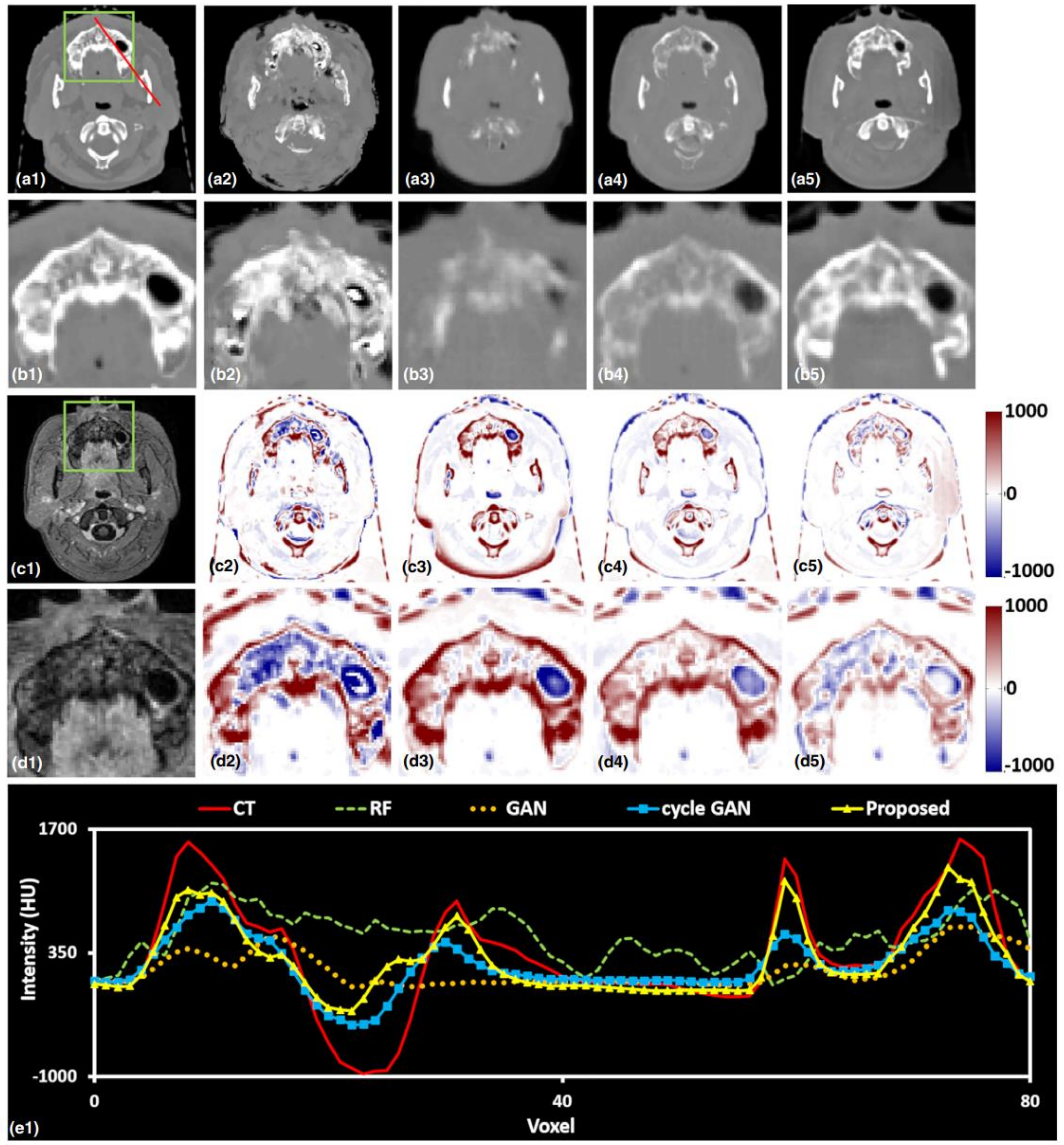
(c) $G_3(I_{LD})$ (d) CAC mask $G_3(I_{LD})$ (g) I_{RD} (h) CAC mask I_{RD}

J. M. Wolterink et. al, IEEE Trans. Medical Imaging(2017)

▪ MRI-CT Convsersion

- MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks

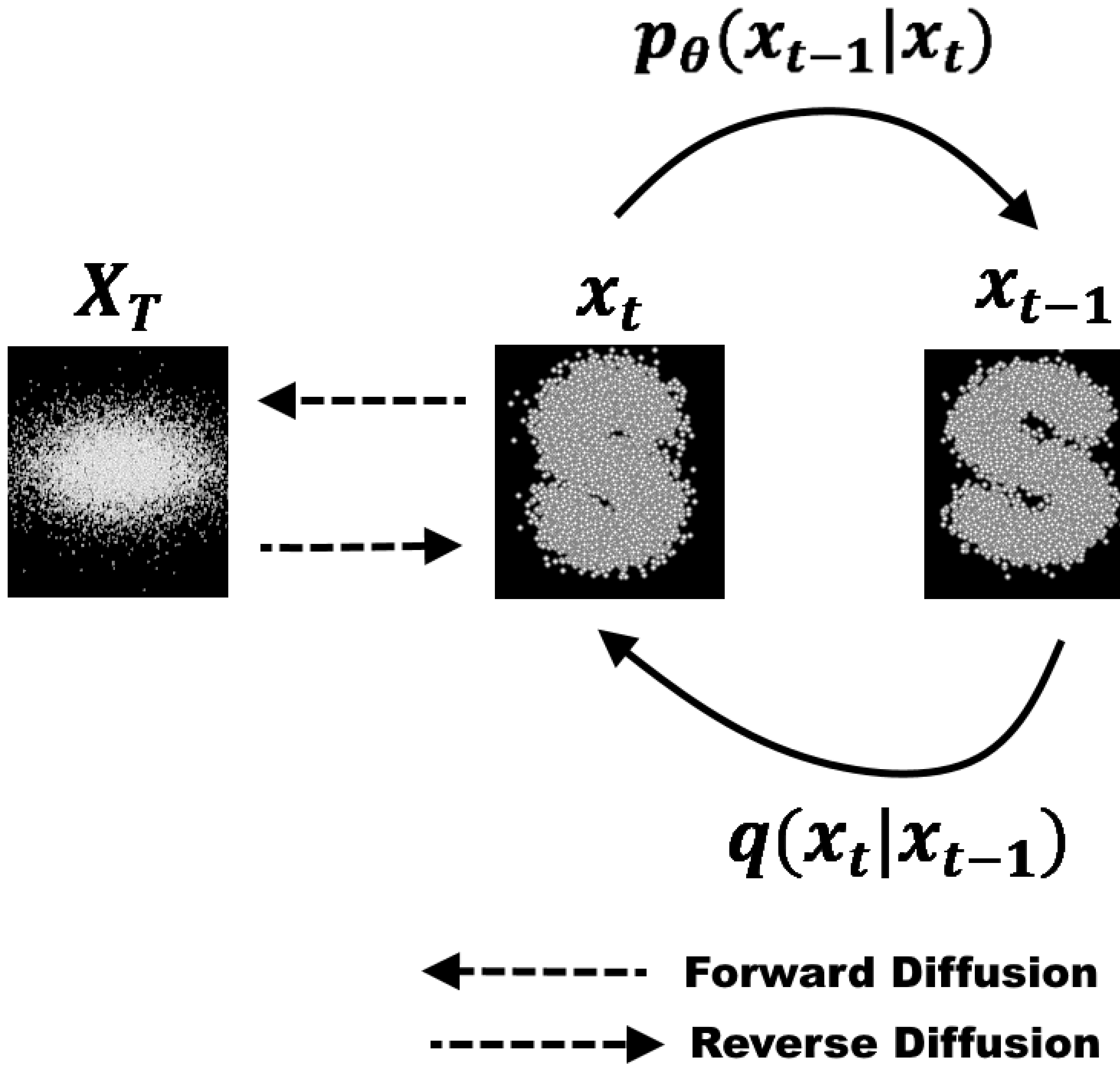


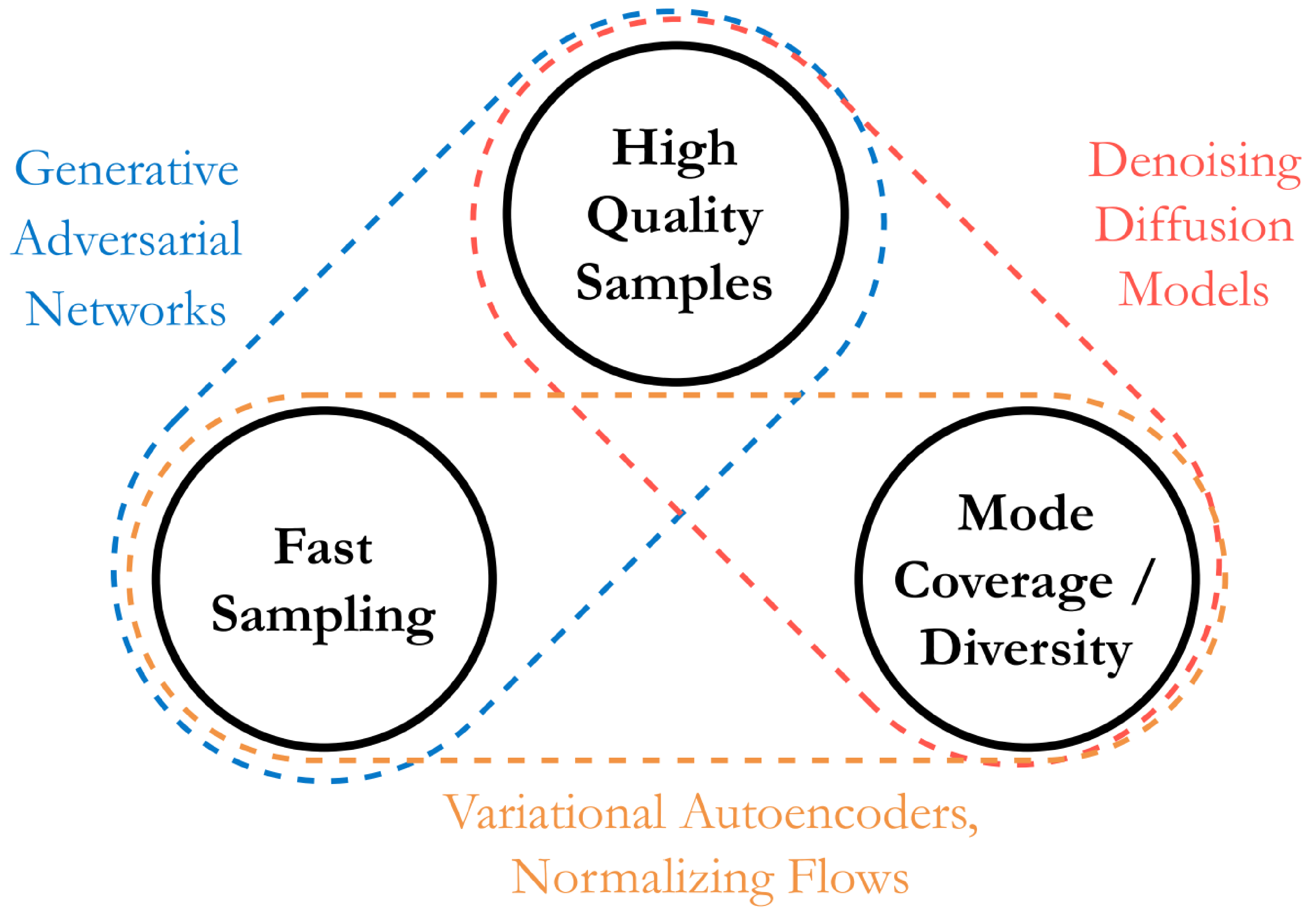


Source : Yang Lei et al., *Medical Physics*(2019)

Image Generation using latent Diffusion Models

- What is diffusion process?





Source :https://jalammar.github.io/illustrated-stable-diffusion/

▪ Generation of Chest X-ray Images from Text Promt

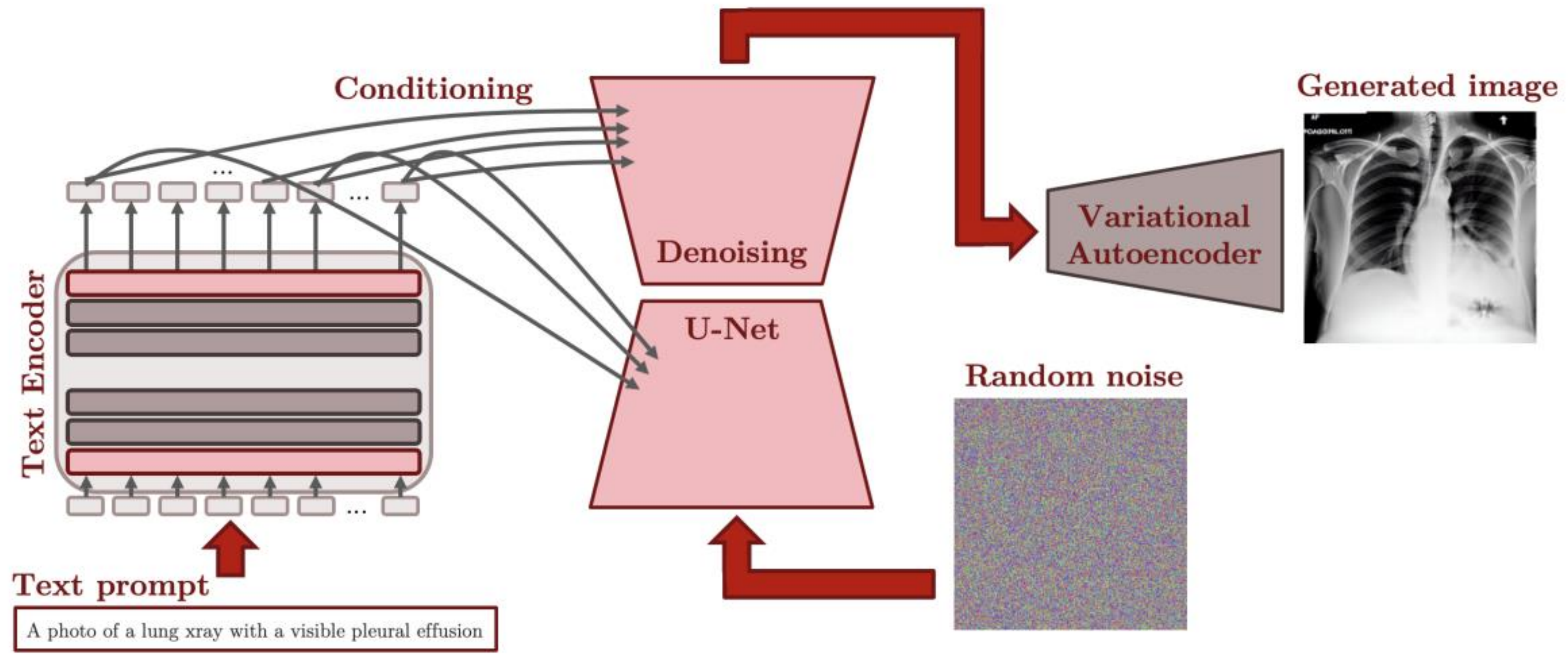
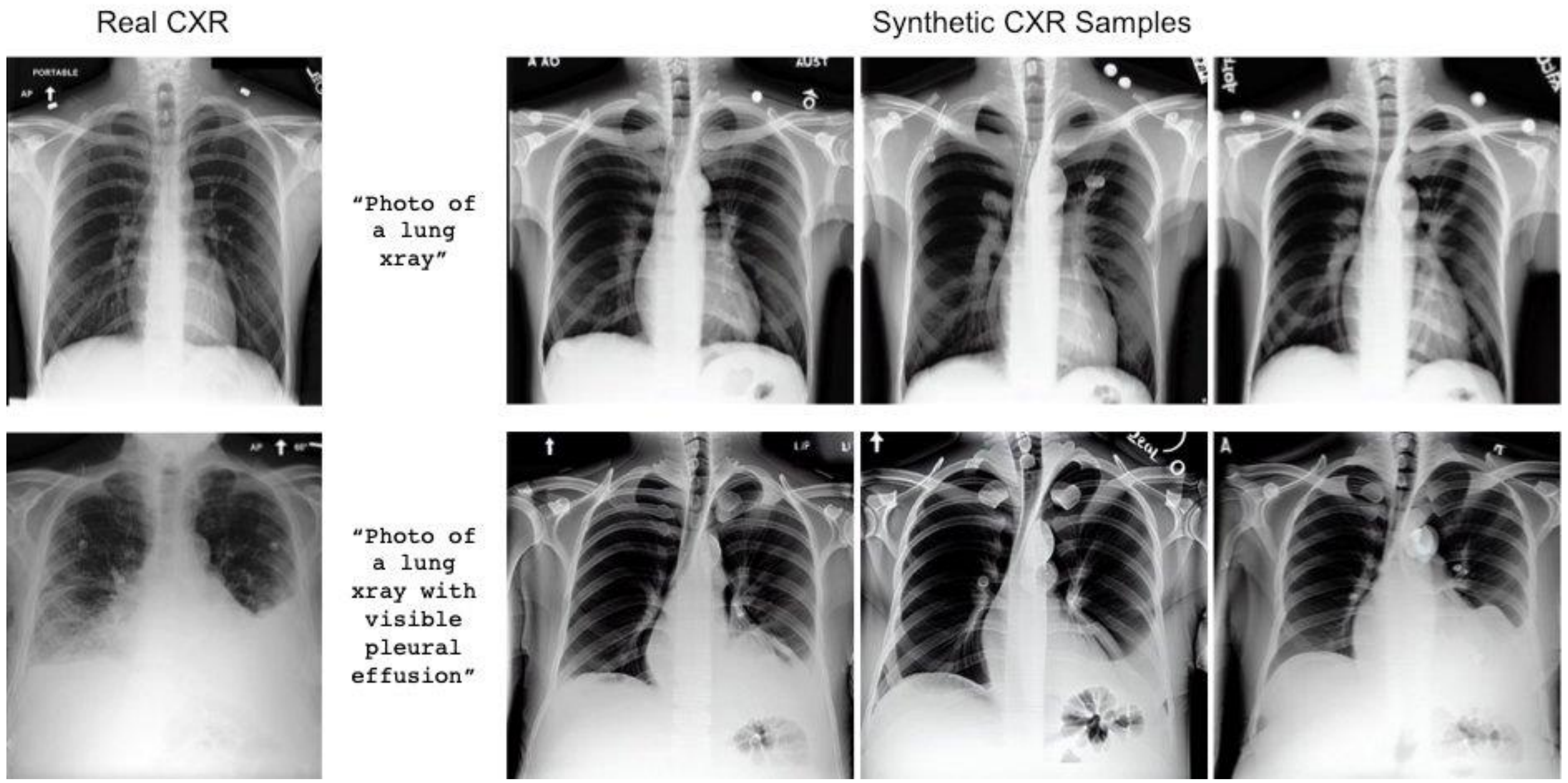


Figure 2: Stable diffusion architecture, run in the radiology setting to generate synthetic radiology images.



Source : P. Sanchez et al.(2022), P. Chambon et al.(2022)

Applications of DDPM to Medical Image Analysis

Anomaly Detection

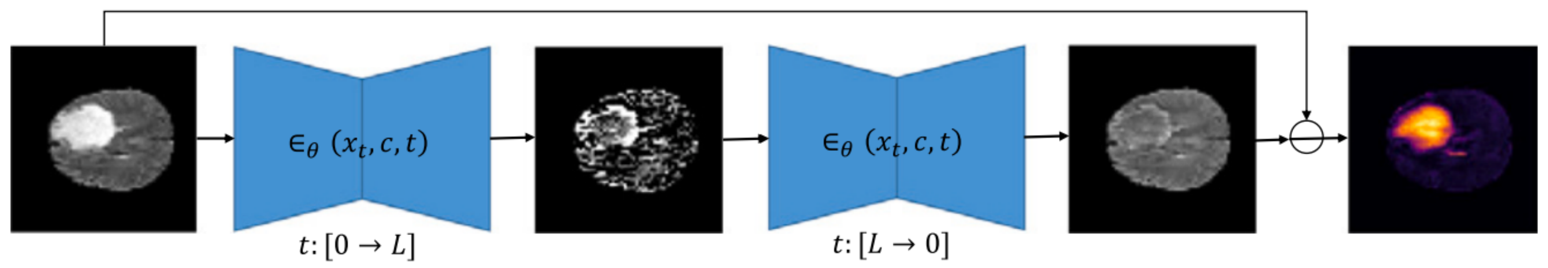


Fig. 14. An overview of CDPM (Sanchez et al., 2022). Iteratively applying diffusion models using an unconditional model ($c = \emptyset$) encodes the input image into a latent space. Then, reversing the diffusion process from the latent space decodes a healthy state image. The decoding process is guided by conditioning it on the healthy state and $\emptyset$. The anomaly heatmap is generated by subtracting the input image from the generated counterfactual.

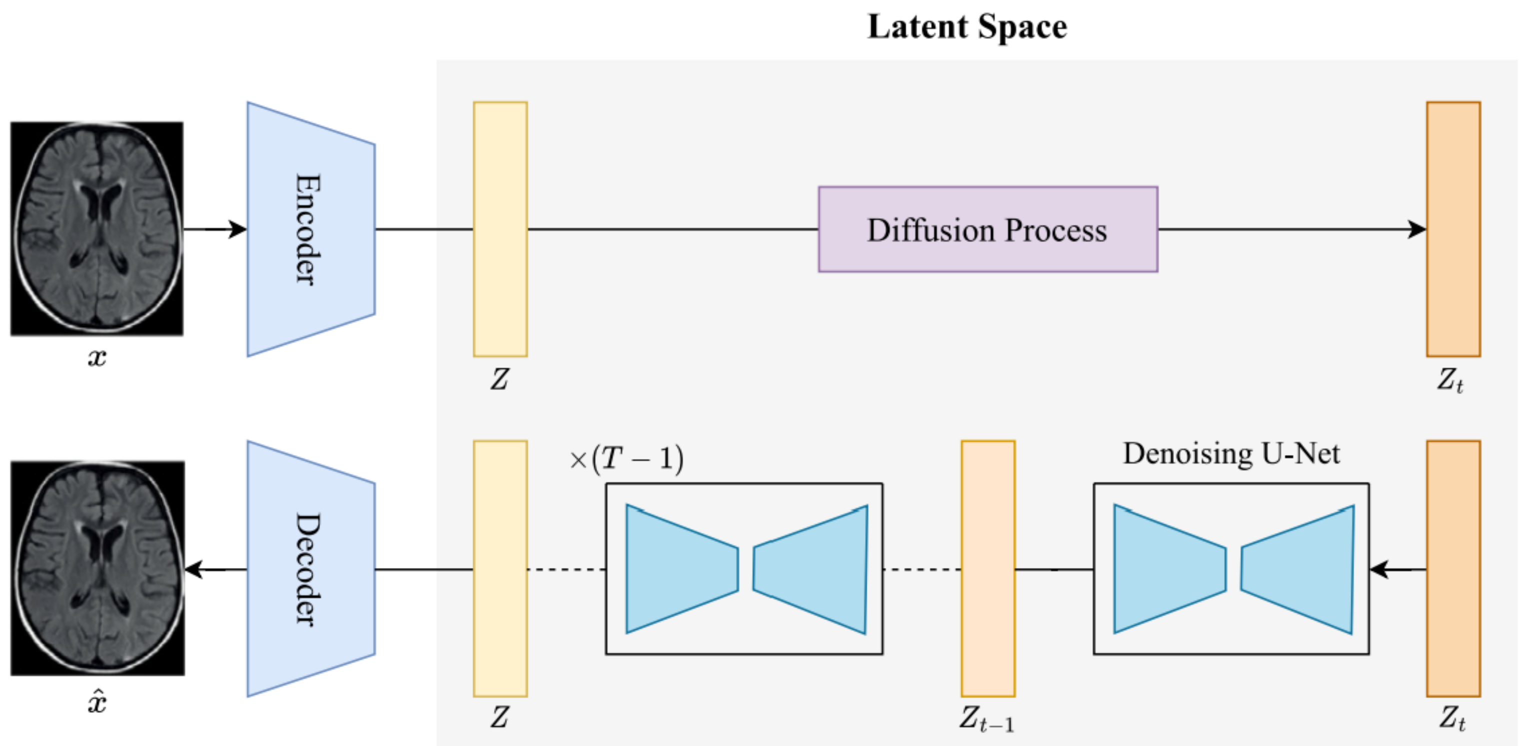


Fig. 15. An overview of BAnoDDPM (Pinaya et al., 2022a). An autoencoder compresses the input image into a latent code, further enhanced by applying diffusion and reverse processes, and decodes into the pixel space.

Generation of Temporal Image

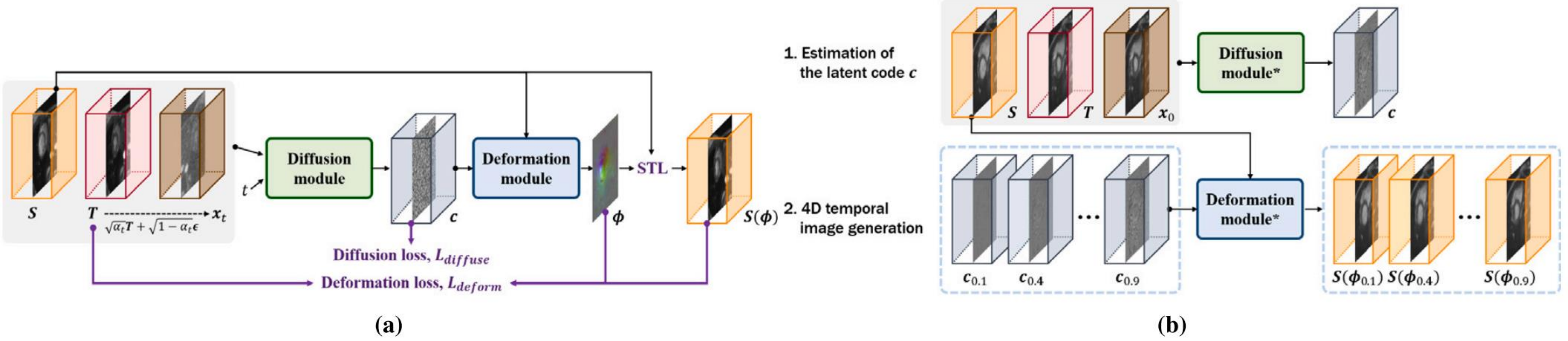
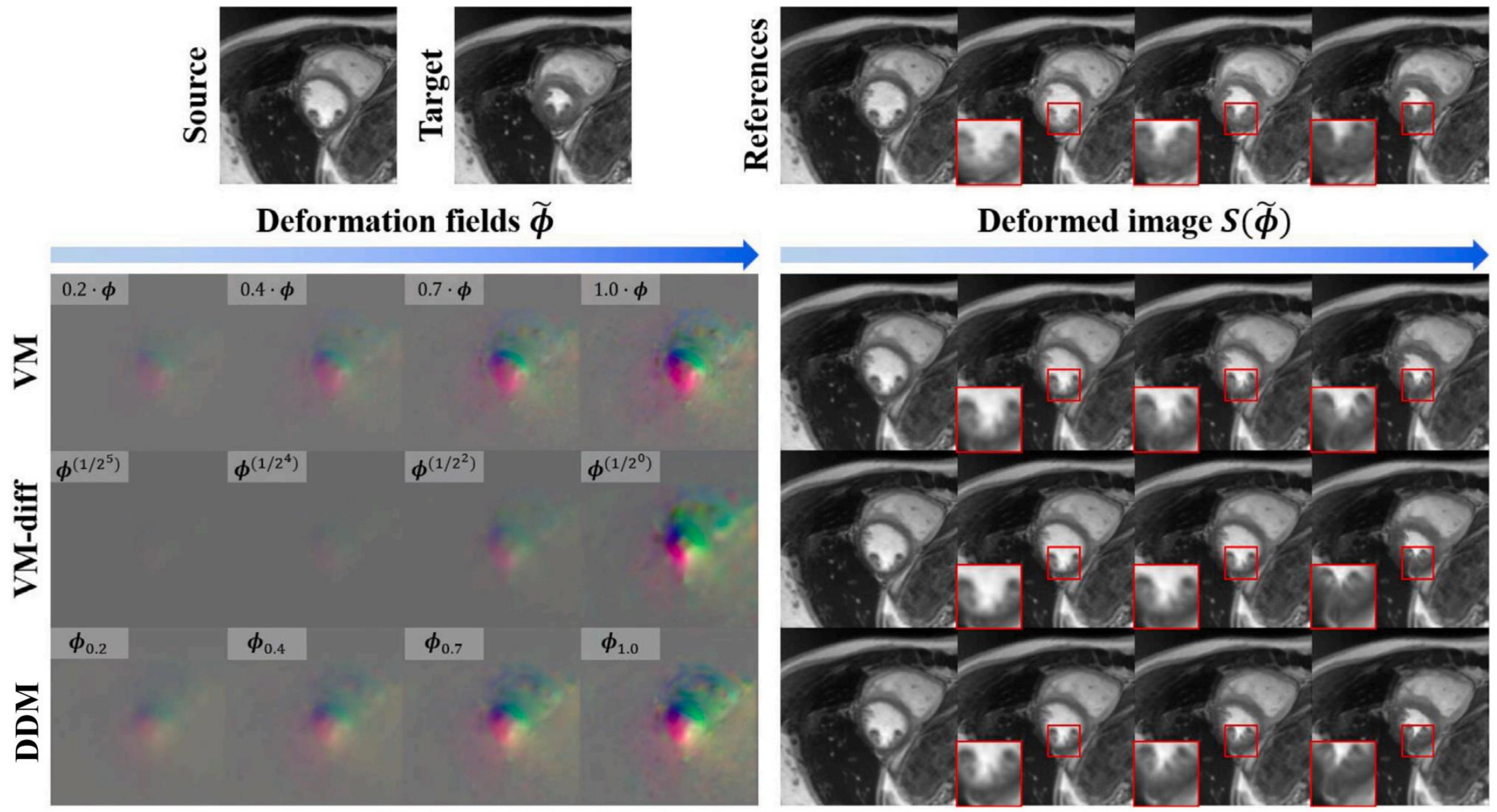


Fig. 12. **(a)** demonstrates the DDM (Kim and Ye, 2022) training phase and **(b)** the inference phase.

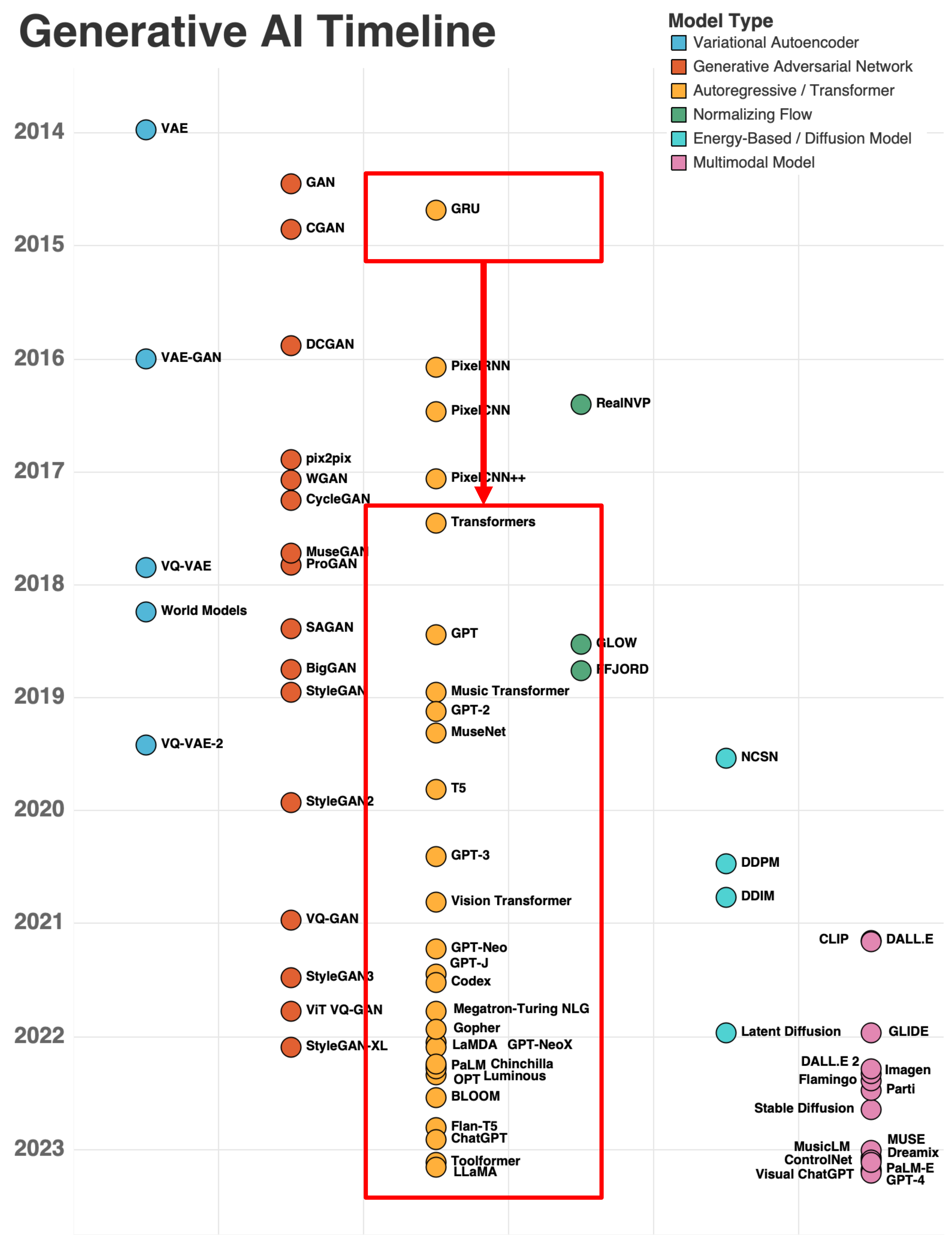


Source : A. Kazerouni et al., *Medical Image Analysis*(2023)

▪ Deep Learning Algorithms for Generative AI

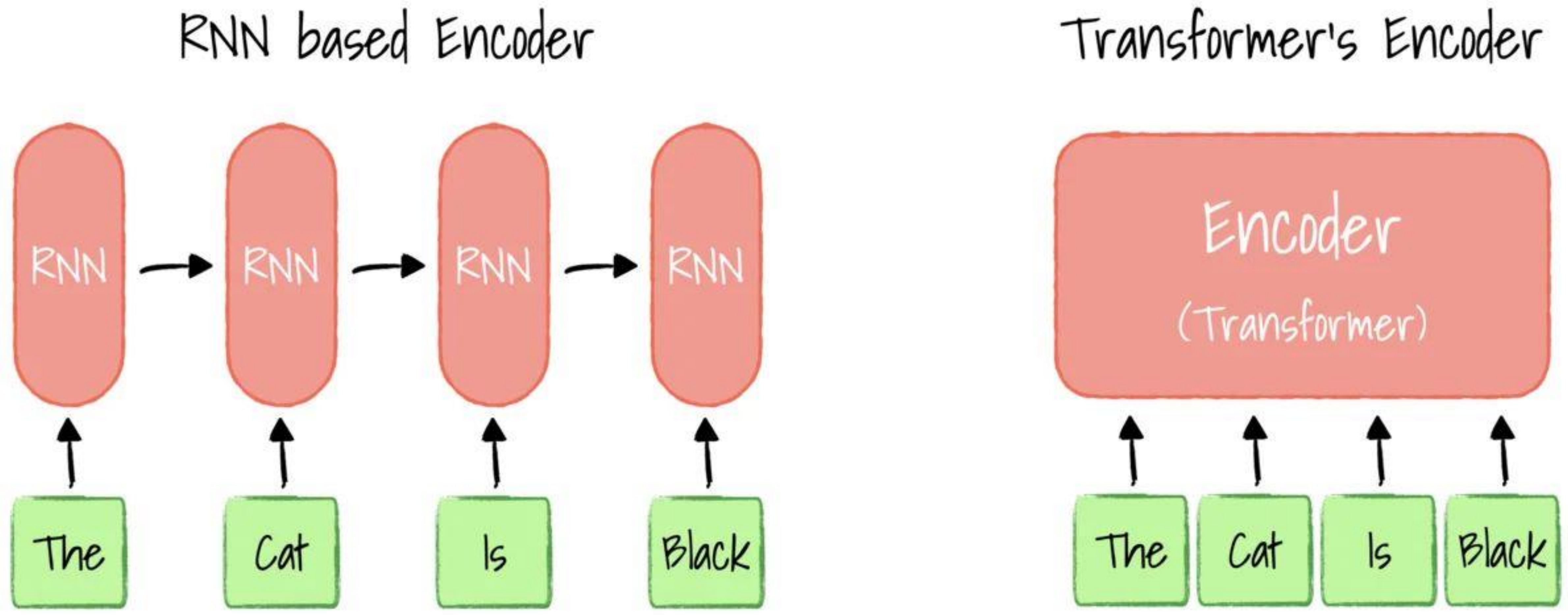
- **Recurrent Neural Networks(RNNs)**
  - A type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (NLP), speech recognition.

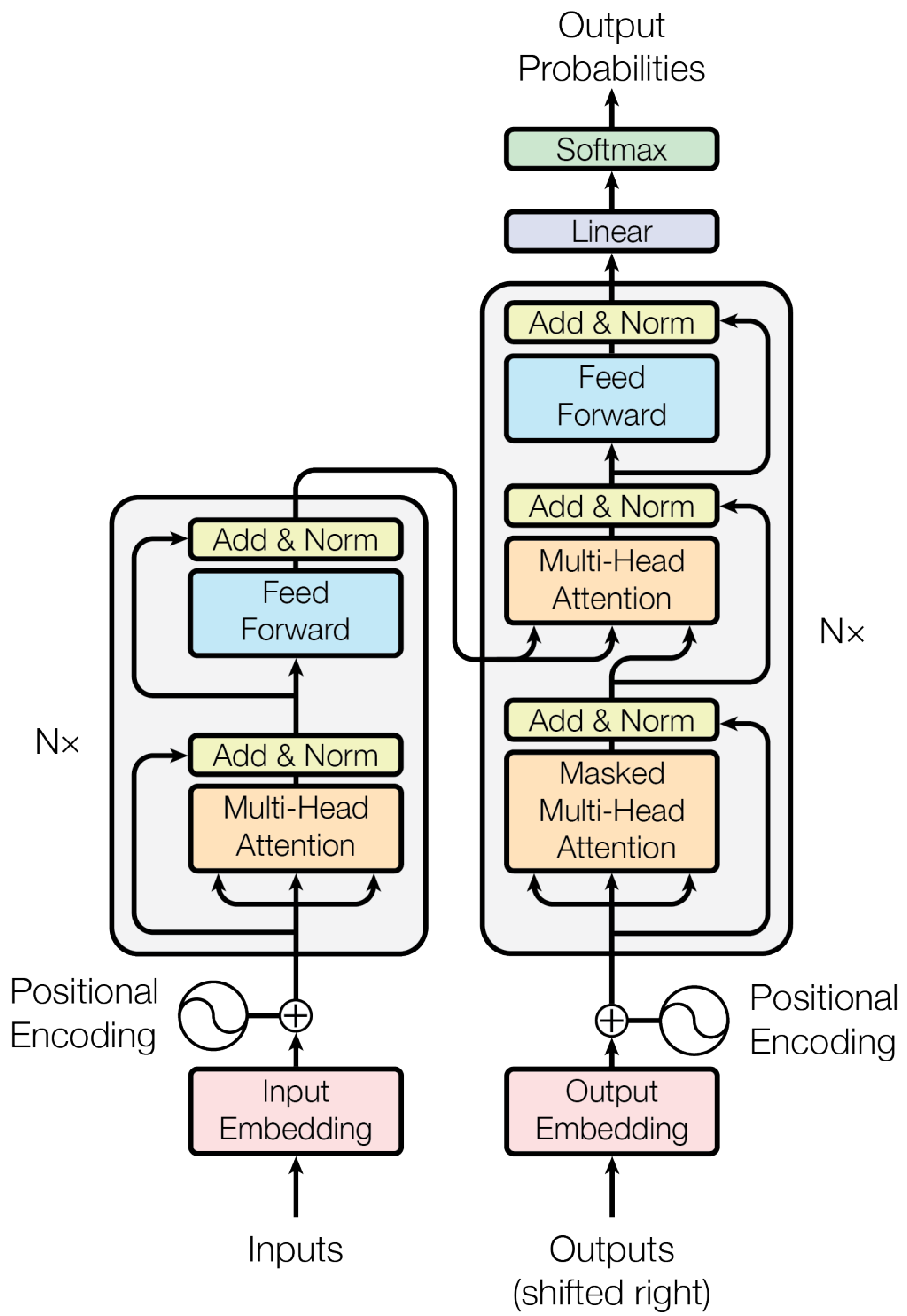
- **Transformers and GPT**
  - A deep learning architecture trained to generate sequences using self-attention mechanism. The autoencoding encoder part of Transformer is called BERT while the autoregressive decoder part is called GPT.



https://github.com/davidADSP/Generative_Deep_Learning_2nd_Edition

- **Recurrent Neural Network vs Transformers**

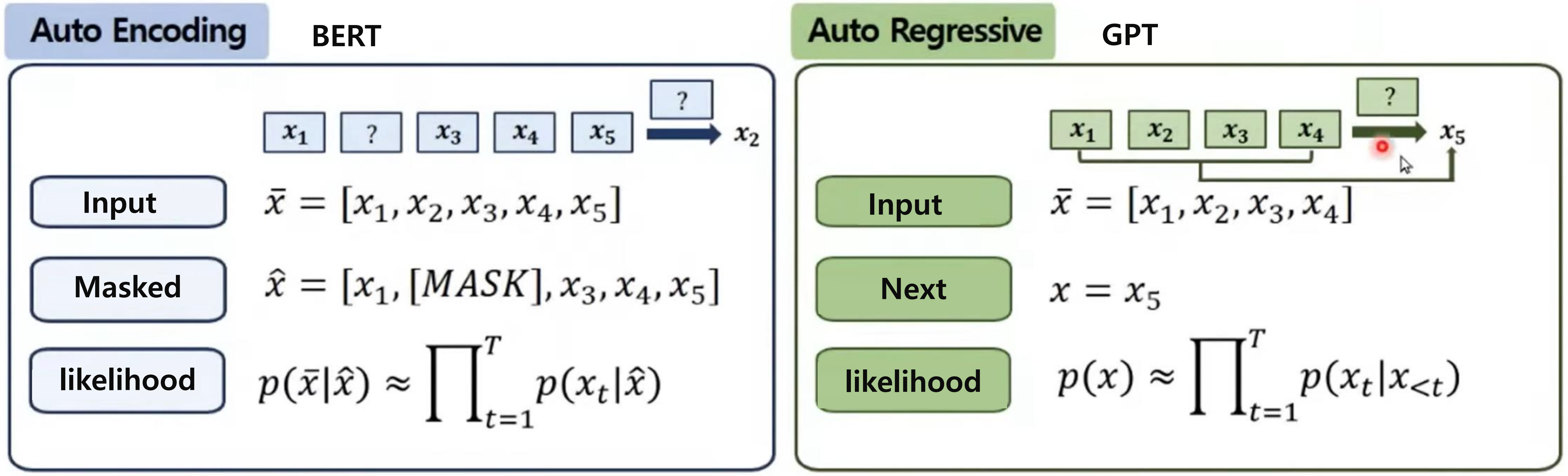


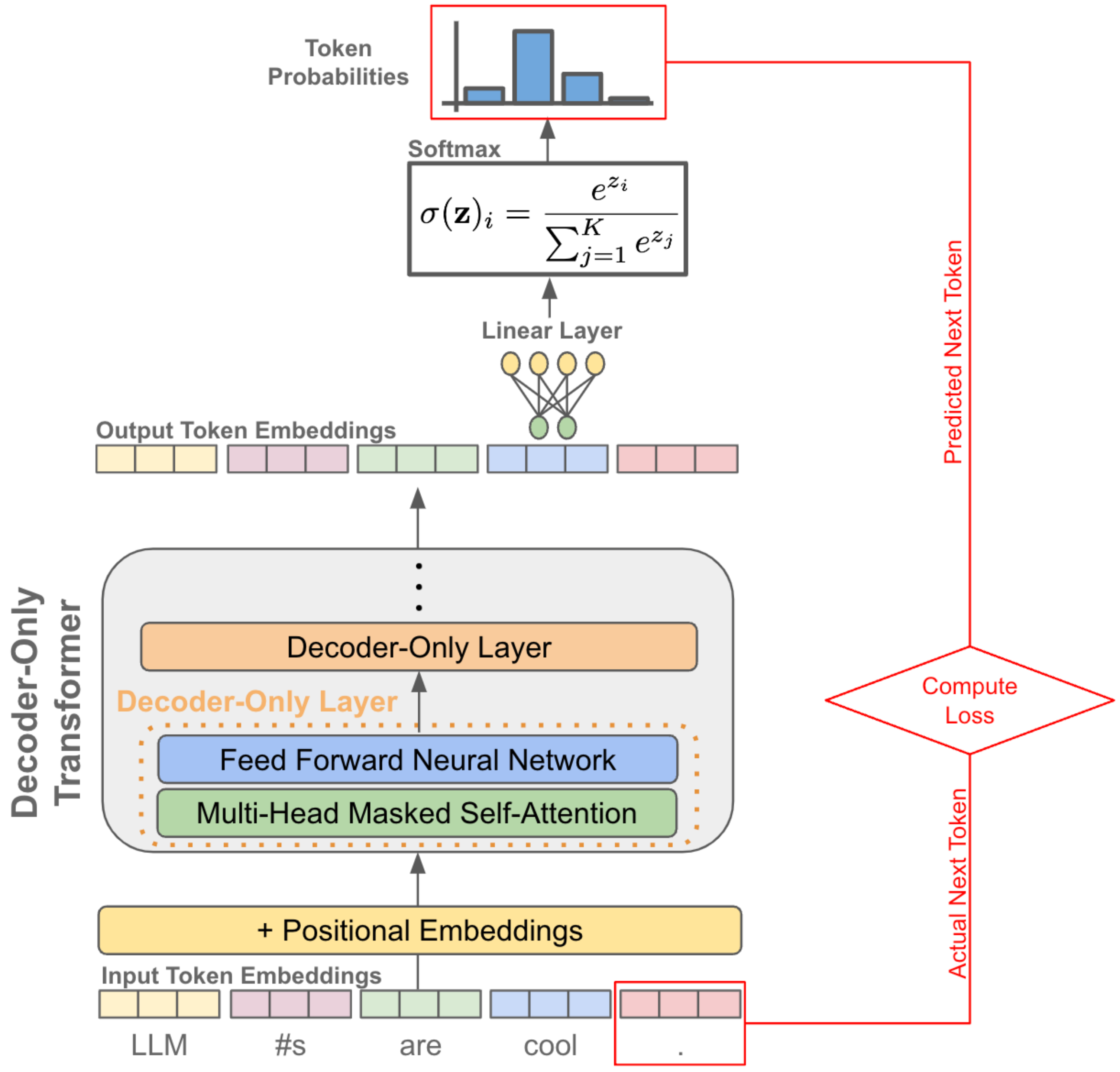


https://jinglescode.github.io/2020/05/27/illustrated-guide-transformer/

Transformer Architecture

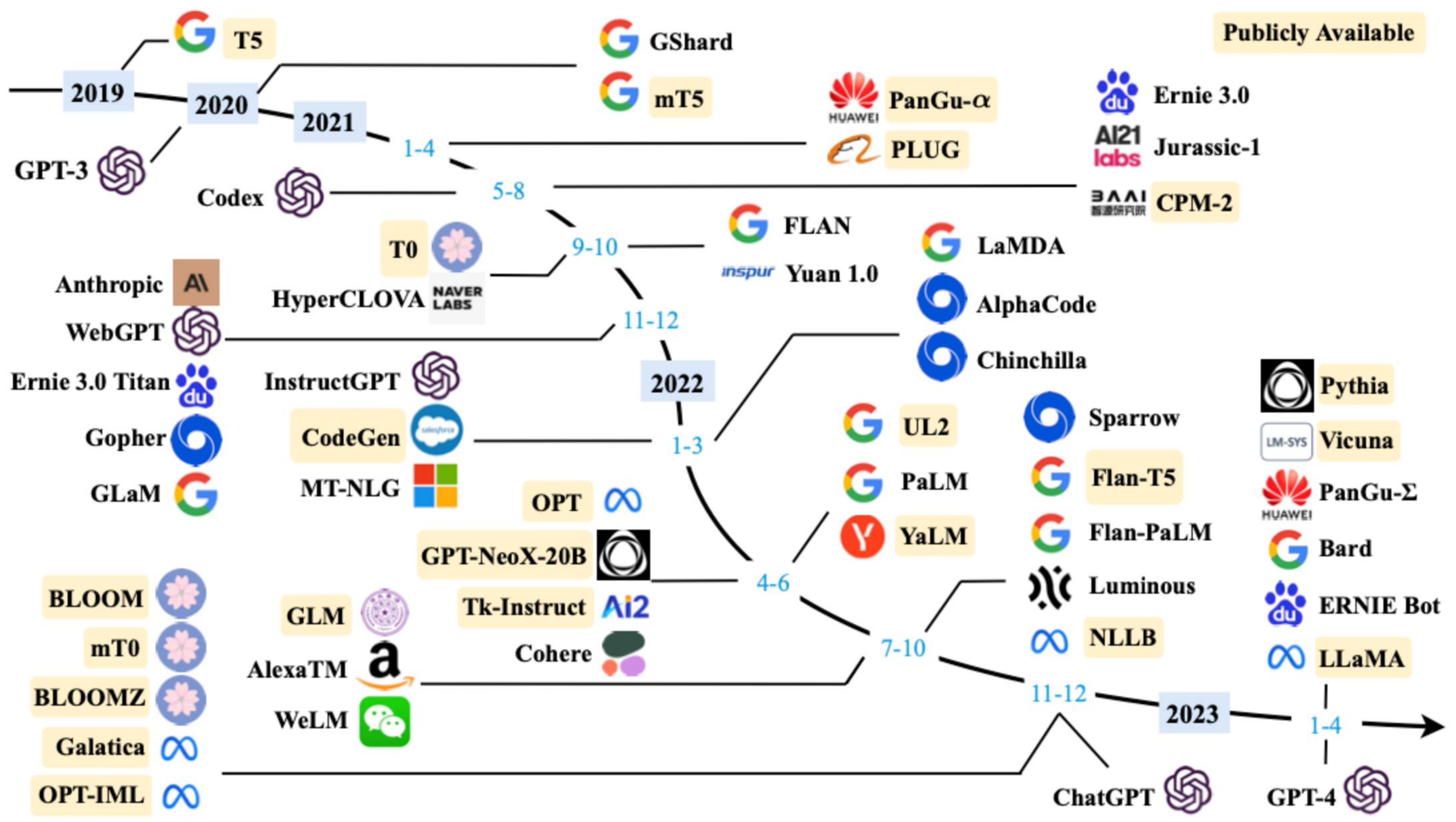
- BERT is a Transformer encoder, which means that, for each position in the input, the output at the same position is the same token (or the [MASK] token for masked tokens), that is the inputs and output positions of each token are the same. Models with only an encoder stack like BERT generate all its outputs at once.
- GPT is an autoregressive transformer decoder, which means that each token is predicted and conditioned on the previous token. This makes these models really good at tasks like language generation, but not good at classification. These models can be trained with unlabeled large text corpora from books or web articles.

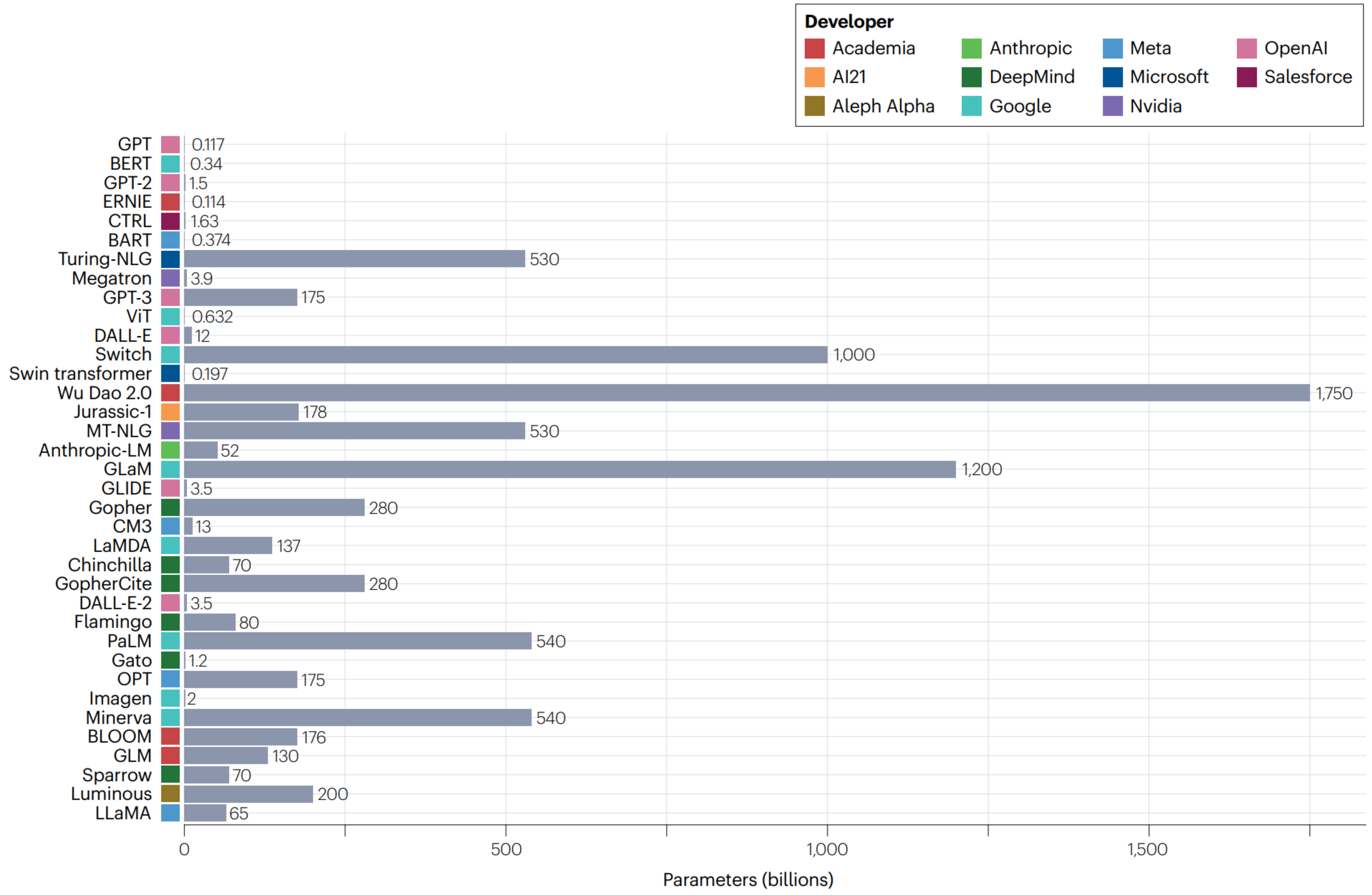




https://jinglescode.github.io/2020/05/27/illustrated-guide-transformer/

▪ Timeline and Model Size of Large Language Models

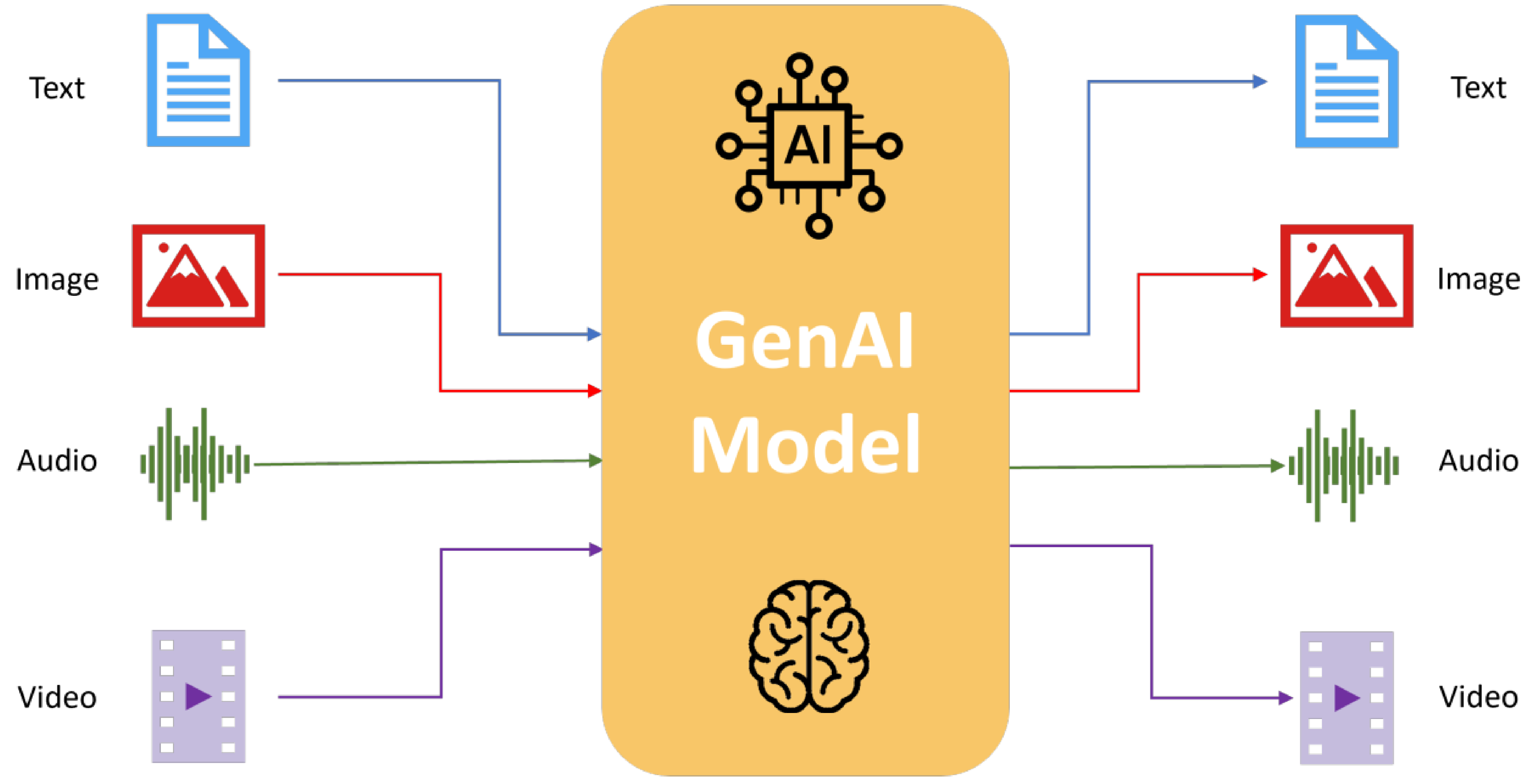


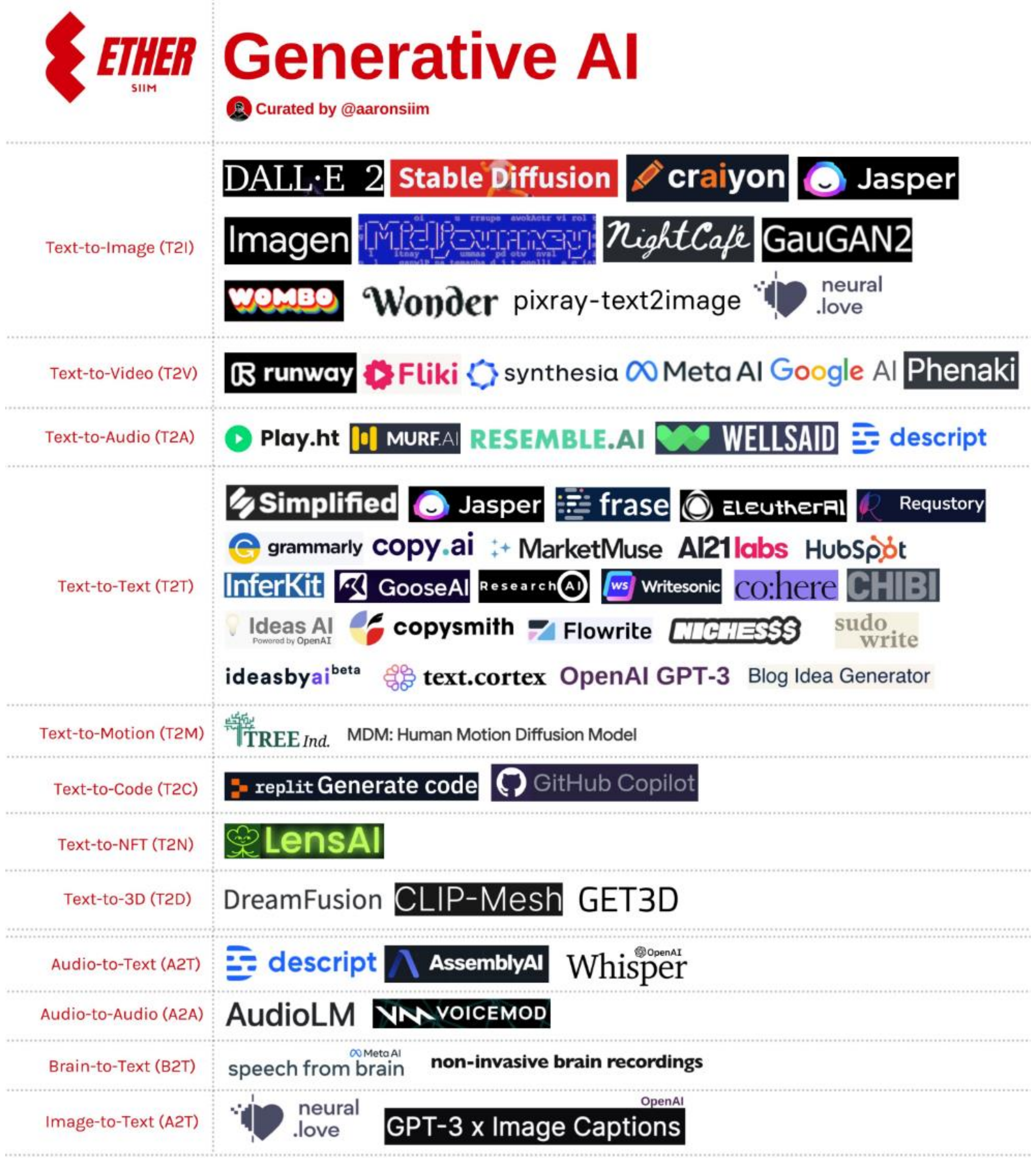


https://velog.io/@easter423/GPT-3-vs-GPT-3.5-vs-ChatGPT, A. J. Thirunavukarasu et al., *Nature Medicine* (2023)

Multimodal AI

- A single model that can digest and generate cross-domain dataset

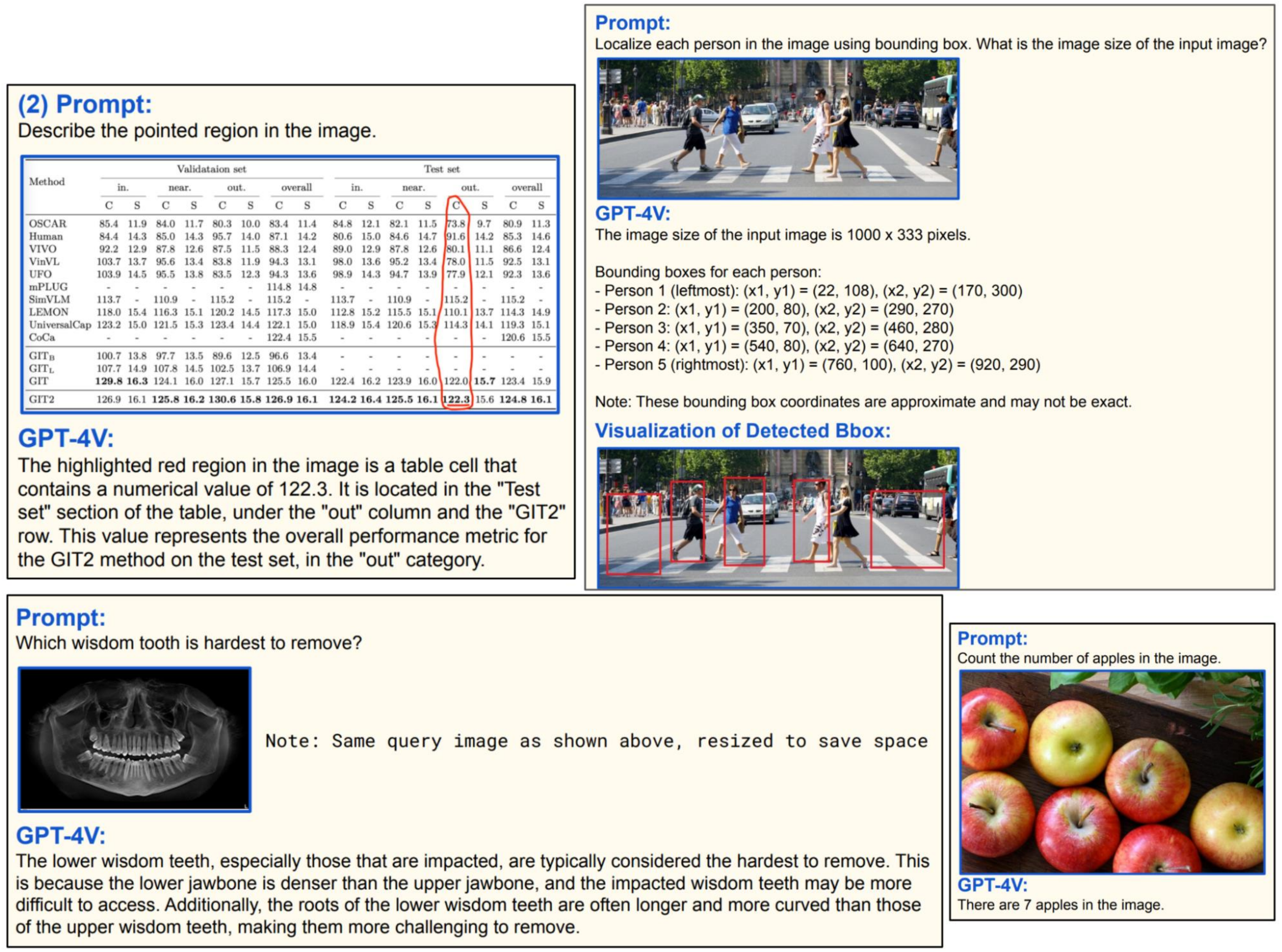


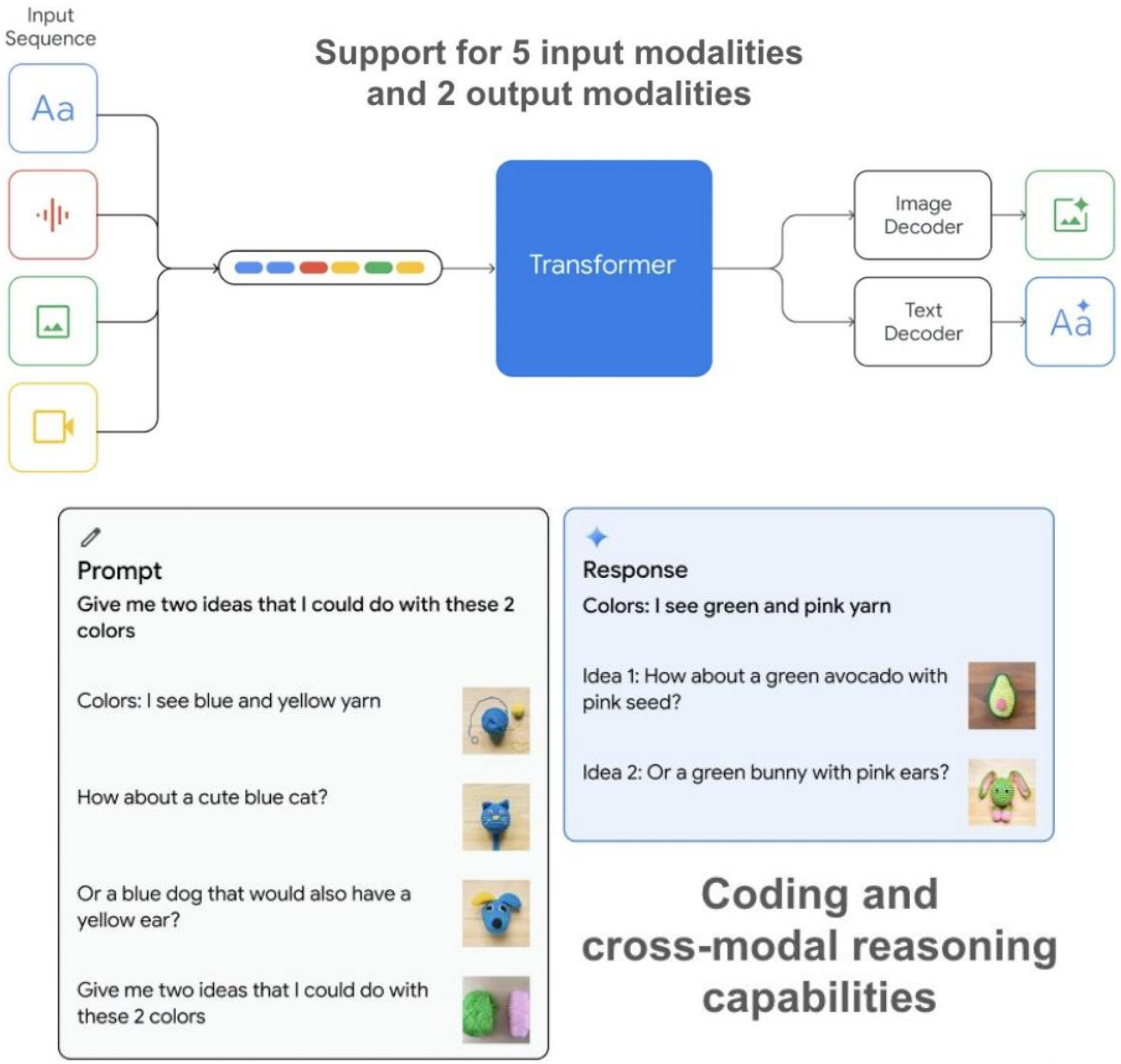


https://www.linkedin.com/pulse/multimodal-generative-ai-tarun-sharma-zzf9c/, @aaronsim

Multimodal AI

- A single model that can digest and generate cross-domain dataset





OpenAI(2023), Google(2023)

02

Medical Applications of Generative AI

▪ Performance of ChatGPT on a Radiology Board-style Examination

Radiology

ORIGINAL RESEARCH • HEALTH POLICY AI

GPT-4 in Radiology: Improvements in Advanced Reasoning

Rajesh Bhayana, MD, FRCPC • Robert R. Bleakney, MD • Satheesh Krishna, MD

From University Medical Imaging Toronto, Joint Department of Medical Imaging, University Health Network, Mount Sinai Hospital and Women's College Hospital, University of Toronto, 200 Elizabeth St, Peter Mulk Building, 1st Fl, Toronto, ON, Canada M5G 24C. Received April 17, 2023; revision requested and received April 24; accepted April 26. **Address correspondence to** R.B. (email: *rajesh.bhayana@uhn.ca*).

Conflicts of interest are listed at the end of this article.

See also the article by Bhayana et al and the editorial by Lourenco et al in this issue.

Supplemental material is available for this article.

Radiology 2023; 307(5):e230987 • https://doi.org/10.1148/radiol.230987 • Content codes: AI HP •

© RSNA, 2023

ChatGPT is a powerful neural network model that belongs to the generative pretrained transformer (GPT) family of large language models (LLMs). Despite being created primarily for humanlike conversations, ChatGPT has shown remarkable versatility and has the potential to revolutionize many industries. It was recently named the fastest growing application in history (1). ChatGPT based on GPT-3.5 nearly passed a text-based radiology examination, performing well on knowledge recall but struggling with higher-order thinking (2). OpenAI's latest LLM, GPT-4, was released in March of 2023 in limited form to paid users alongside claims of enhanced advanced reasoning capabilities (3). GPT-4 demonstrated remarkable improvements over GPT-3.5 on professional and academic benchmarks, including the uniform bar examination (90th vs 10th percentile) and U.S. Medical Licensing Examination (>30% improvement) (4,5).

Despite improved performance on various general professional benchmarks, whether GPT-4's enhanced advanced reasoning capabilities translate to improved performance in radiology, where the context of specific technical language is crucial, remains uncertain. The purpose of this exploratory study was to evaluate the performance of GPT-4 on a radiology board–style examination without images and compare it with that of GPT-3.5.

Materials and Methods

In this prospective study, the performance of GPT-4 was assessed on the same 150 multiple-choice text-based questions used to benchmark GPT-3.5, with the selection process and categorization described previously (2). Questions matched the style, content, and difficulty of the Canadian Royal College and American Board of Radiology examinations. GPT-4 performance was assessed overall, by question type, and by topic. GPT-4's

Performance of GPT-4 and GPT-3.5 on Radiology Board–style Multiple-Choice Questions without Images, Stratified by Question Type and Topic

Parameter	No. of Questions	No. of Correct Responses: GPT-3.5	No. of Correct Responses: GPT-4	*P* Value
All questions	150	104 (69.3)	121 (80.7)	.02
Question type				
Lower-order thinking	61	51 (84)	49 (80)	.64
Higher-order thinking	89	53 (60)	72 (81)	.002
Description of imaging findings	46	28 (61)	39 (85)	.009
Clinical management	18	16 (89)	16 (89)	>.99
Application of concepts	10	3 (30)	9 (90)	.006
Calculation and classification	8	2 (25)	4 (50)	.30
Disease associations	7	4 (57)	4 (57)	>.99
Topic				
Physics	15	6 (40)	13 (87)	.008
Clinical	135	98 (73)	108 (80)	.15
Gastrointestinal	16	10 (63)	13 (81)	.24
Genitourinary	18	14 (78)	14 (78)	>.99
Bone	18	12 (67)	13 (72)	.72
Breast	15	12 (80)	13 (87)	.62
Chest	17	13 (76)	13 (76)	>.99
Cardiovascular	18	13 (72)	13 (72)	>.99
Head and neck	16	11 (69)	15 (94)	.07
Systemic	17	13 (76)	14 (82)	.67

Note.—Numbers in parentheses are percentages.

R

A 64-year-old man undergoes a chest CT scan while in hospital for persistent pneumonia. A 3-cm adrenal nodule is incidentally noted, and a subsequent adrenal washout study is ordered. On the non-contrast imaging series, region of interest (ROI) measurement of the nodule provides a reading of 27 HU. On the contrast-enhanced series that follows, this measurement rises to 88 HU. Delayed images through the adrenal glands at 15 minutes show a reading of 45 HU. What is the absolute washout for this lesion?
A) 31%
B) 49%
C) 59%
D) 70%

The formula for absolute percentage washout (APW) is:

(APW) = [(HU unenhanced - HU delayed)/ (HU unenhanced - HU enhanced)] x 100%

Using the values provided in the question, we can calculate the absolute washout for the lesion as follows:

(APW) = [(27 - 45)/(27 - 88)] x 100%
(APW) = (-18)/(-61) x 100%
(APW) = 0.295 x 100%
(APW) = 29.5%

Therefore, the absolute washout for this lesion is 29.5%, which rounds to 31%, so the correct answer is A) 31%.

To calculate the absolute washout for an adrenal lesion, you can use the following formula:

Absolute washout (%) = [(Enhanced HU - Delayed HU) / (Enhanced HU - Non-contrast HU)] x 100

In this case:
Enhanced HU = 88
Delayed HU = 45
Non-contrast HU = 27

Plugging these values into the formula:

Absolute washout (%) = [(88 - 45) / (88 - 27)] x 100
Absolute washout (%) = [43 / 61] x 100
Absolute washout (%) ≈ 70%

Therefore, the correct answer is D) 70%.

Source : R. Bhayana et al., *Radiology*(2023)

▪ Performance of Various Open-source and Commercial LLMs for Nephrology Questions

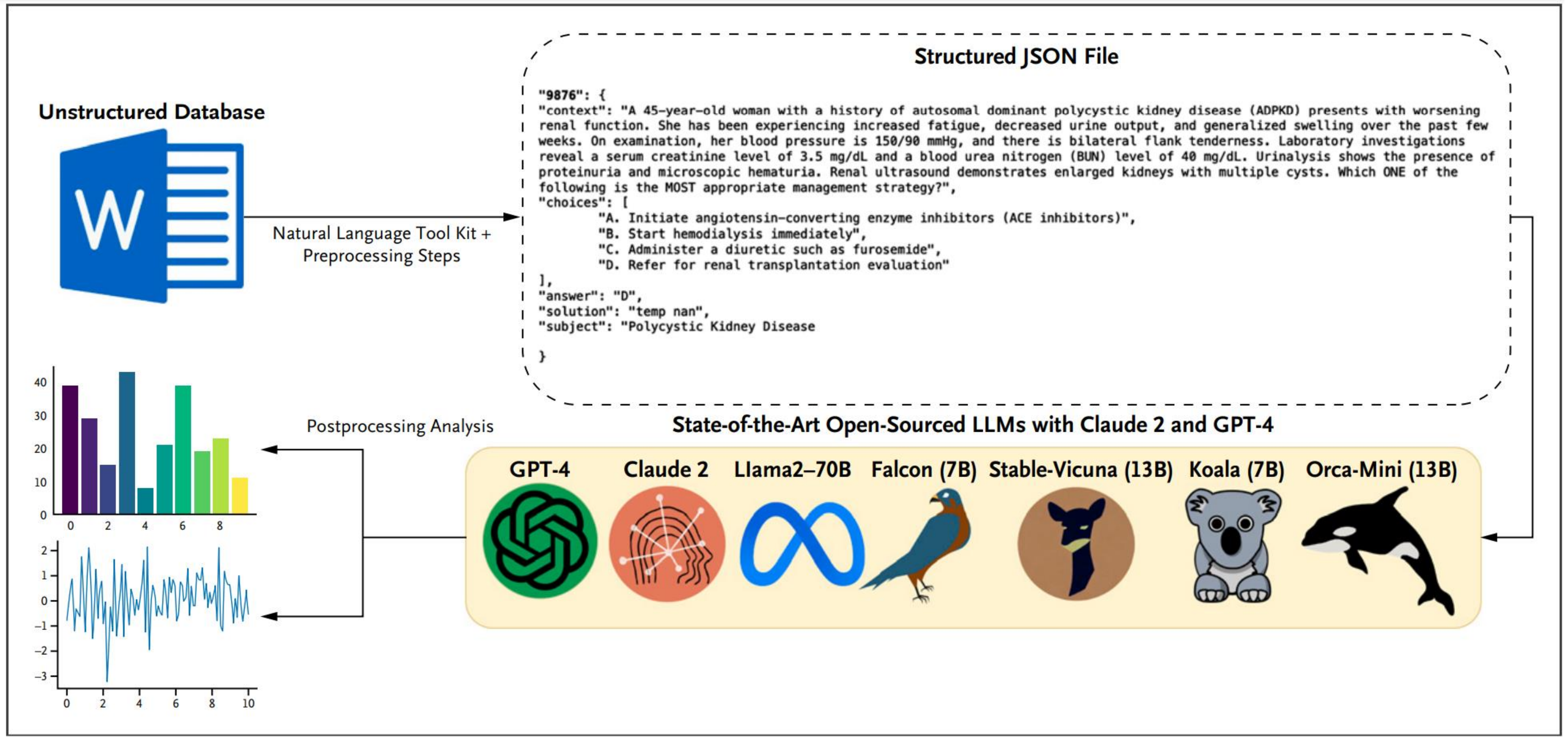
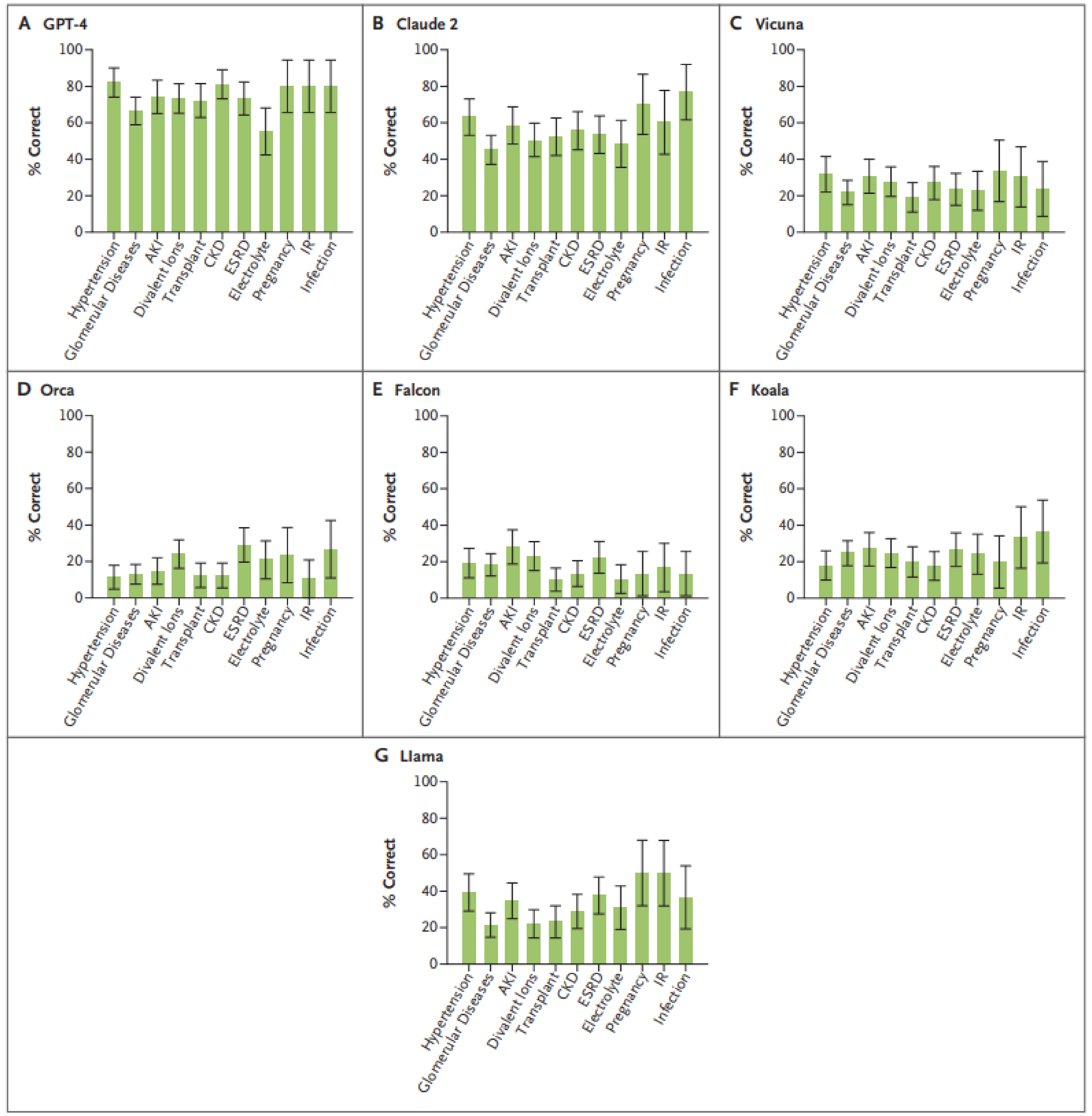


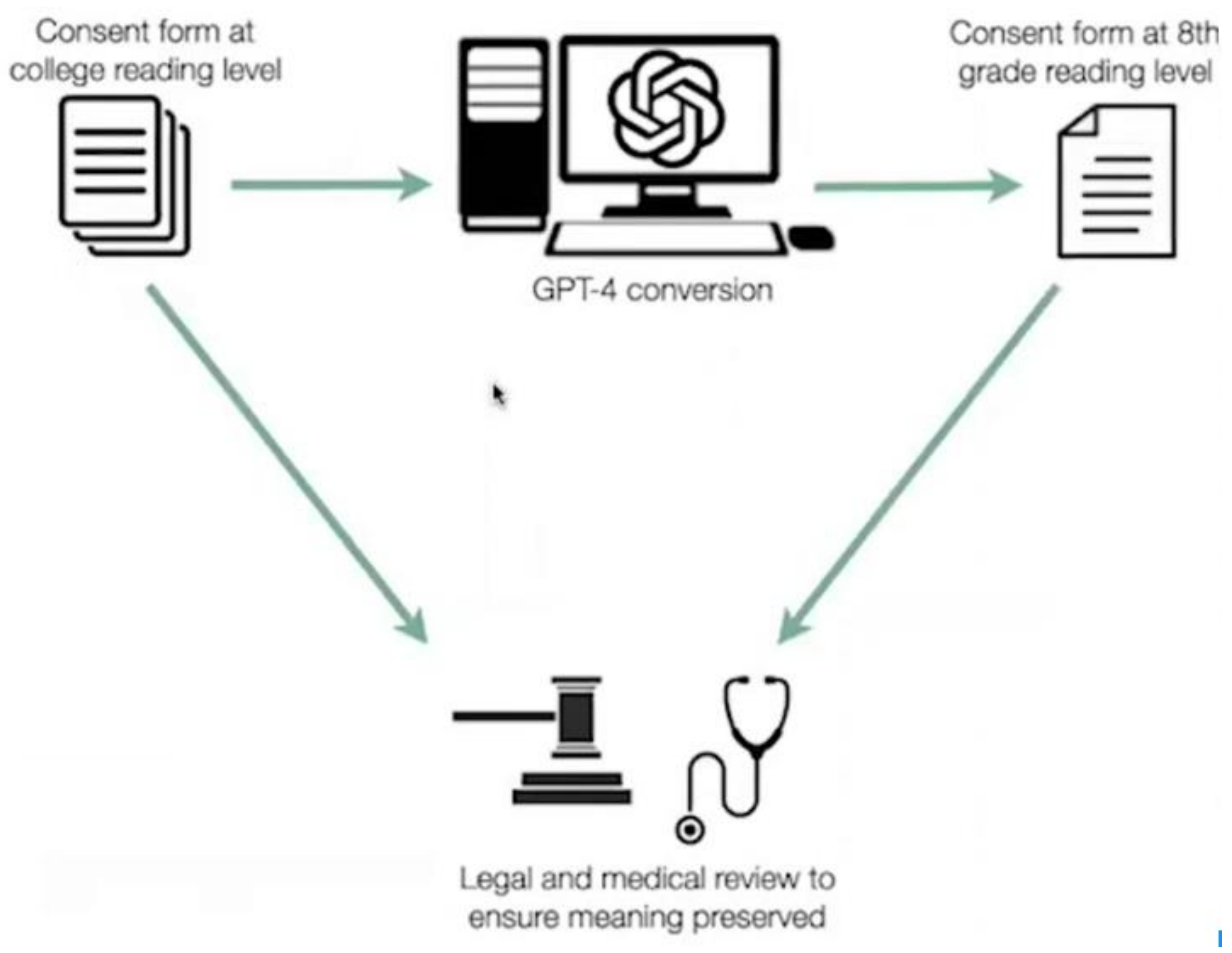
Table 1. Comparison of the Overall Correct Responses among the Large Language Models.*

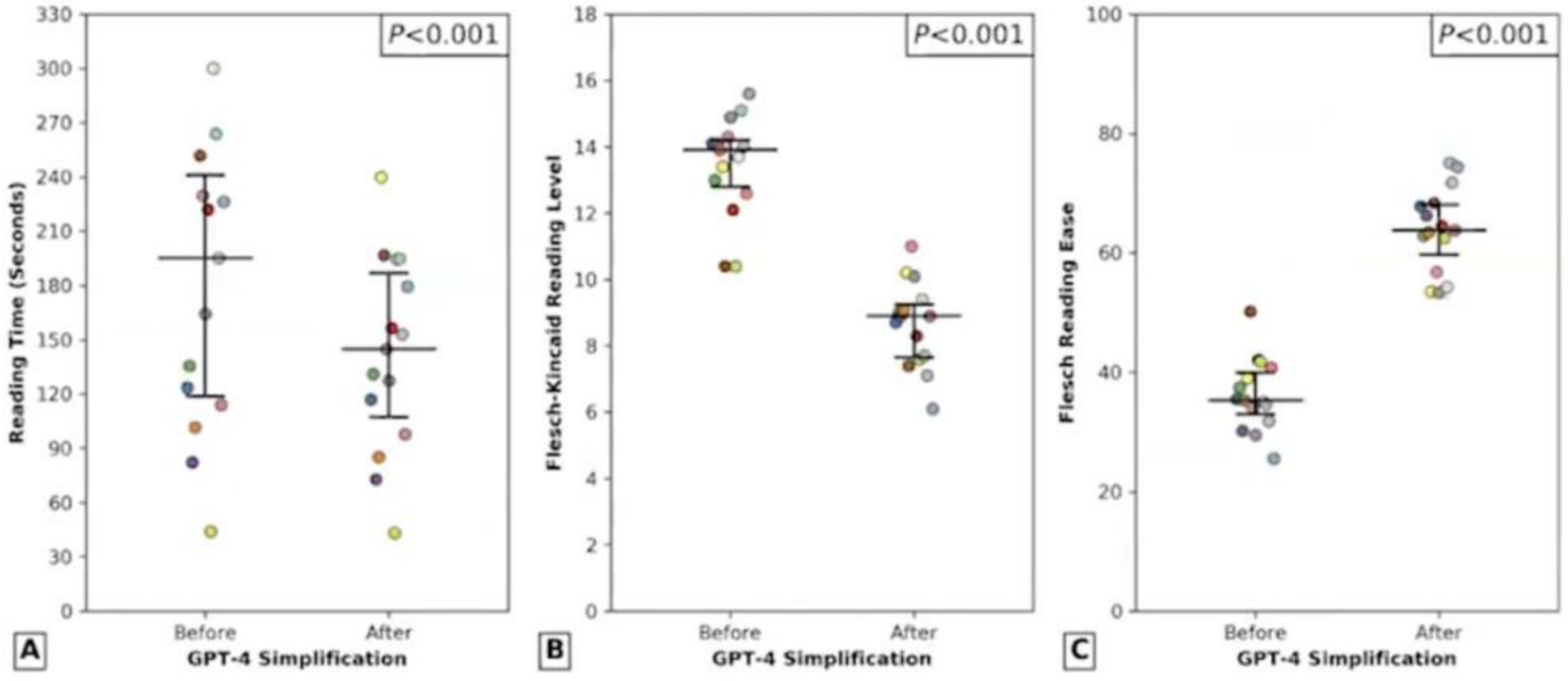
LLM	Total Questions	Number Correct	Percentage Correct	CI
GPT-4	858	629	73.3	70.3–76.3
Claude 2	858	467	54.4	51.1–57.7
Vicuna	858	219	25.5	22.6–28.4
Orca	858	147	17.1	14.6–19.6
Falcon	858	155	18.1	15.5–20.7
Koala	858	204	23.8	21.0–26.6
Llama	858	263	30.6	27.6–33.8

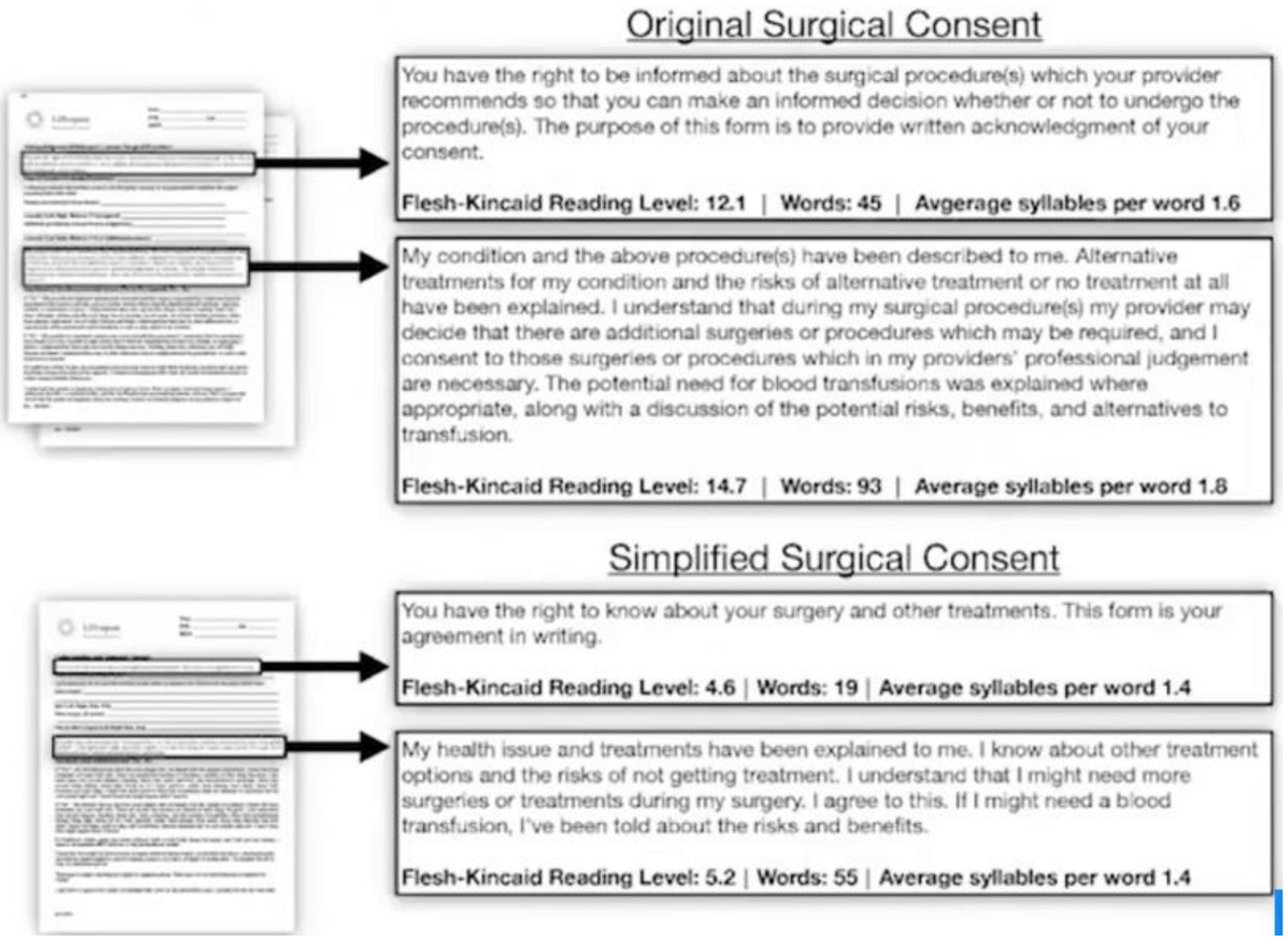


Source : S. Wu et al., *NEJM AI*(2024)

- **Using LLM for Facilitating Truly Informed Consent**

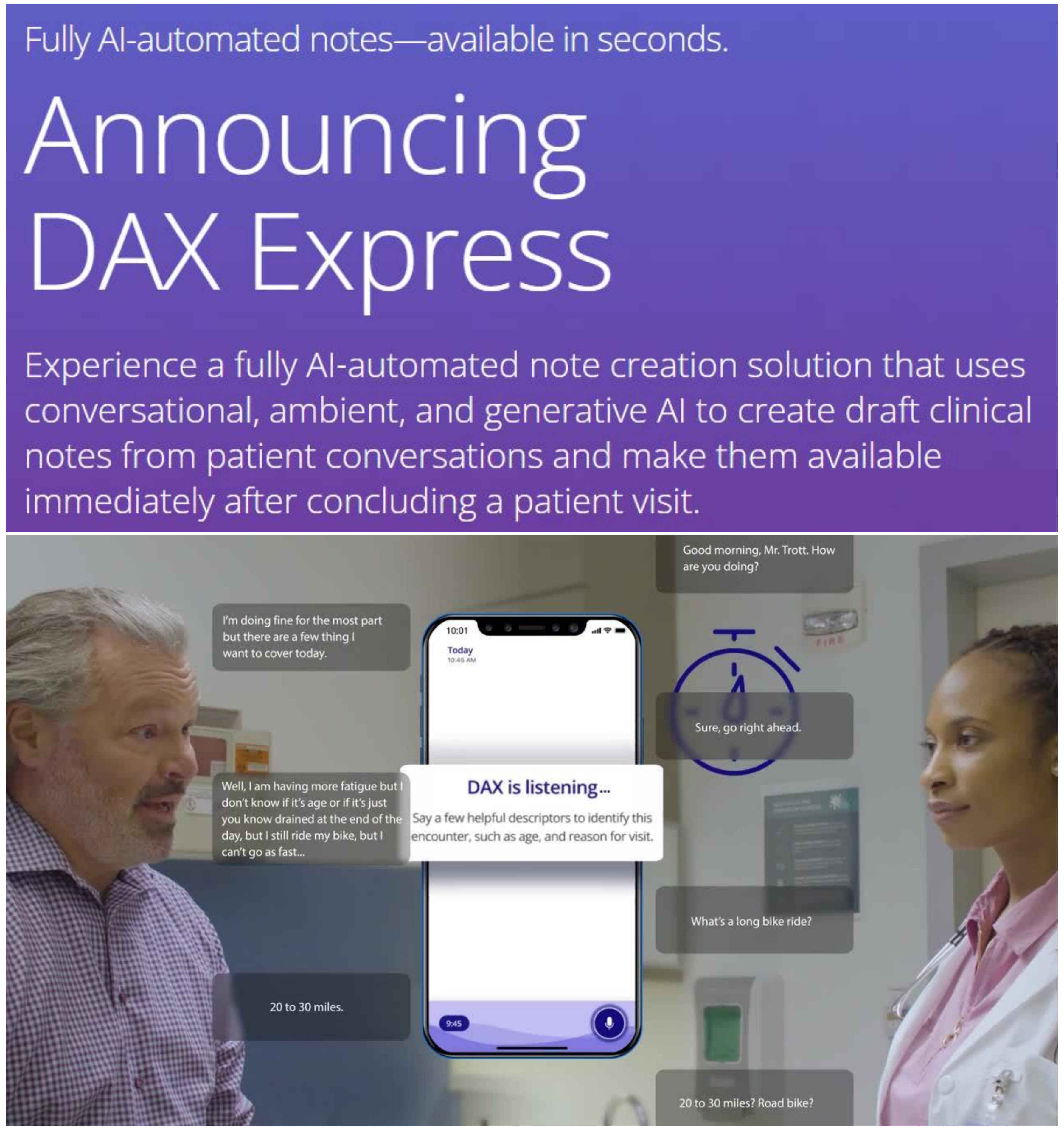






Source : F. N. Mirza et al., *NEJM AI* (2024)

- **Integration of LLMs into EMR/EHR**



A A Request to GPT-4 to Read a Transcript of a Physician–Patient Encounter and Write a Medical Note

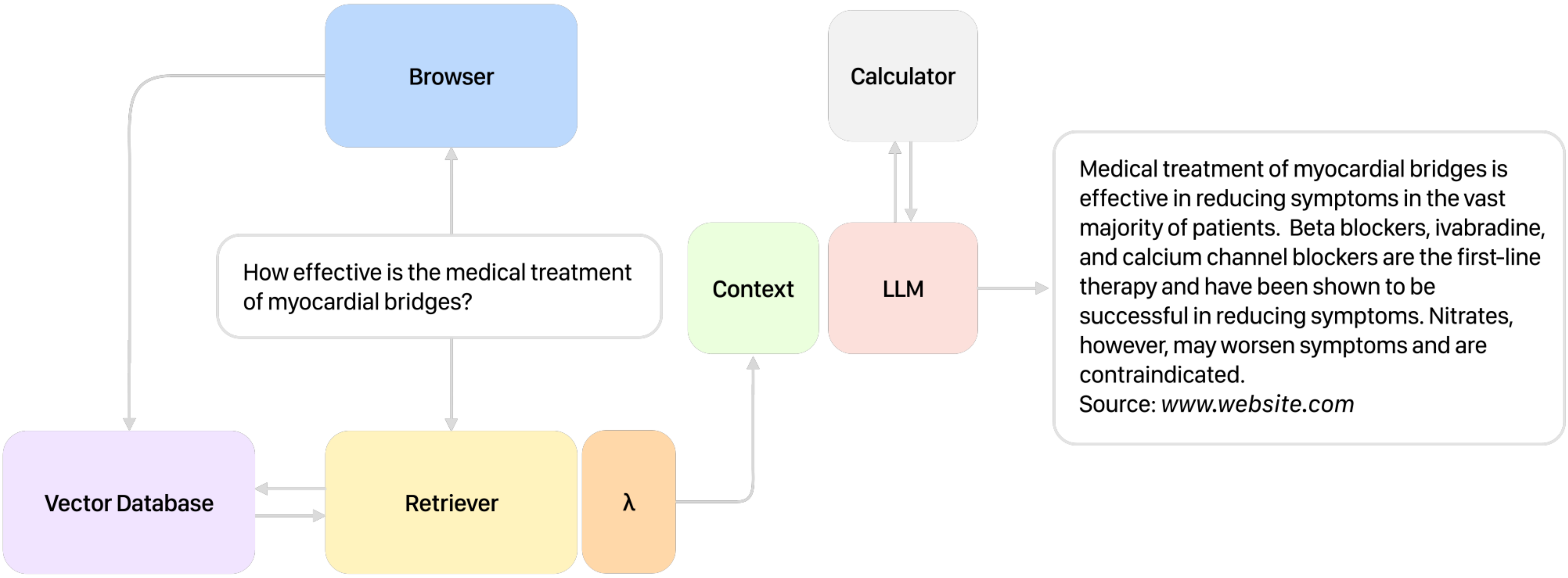
Clinician: Please have a seat, Meg. Thank you for coming in today. Your nutritionist referred you. It seems that she and your mom have some concerns. Can you sit down and we will take your blood pressure and do some vitals?
Patient: I guess. I do need to get back to my dorm to study. I have a track meet coming up also that I am training for. I am runner.
Clinician: How many credits are you taking and how are classes going?
Patient: 21 credits. I am at the top of my class. Could we get this done? I need to get back.
Clinician: How often and far do you run for training now? You are 20, correct?
Patient: Yes. I run nine miles every day.
Clinician: Your BP is 100/50. Your pulse is 52. Meg, how much have you been eating?
Patient: I have been eating fine. I talked to the nutritionist about this earlier.
Clinician: Let's have you stand up and face me and I will back you onto the scale. Eyes on me please. Thank you, and now for a height. Ok looks like 5'5". Go ahead and have a seat.
Patient: How much? Can I please see what the scale says? I'm fat.
Clinician: Please sit up and I will listen to your heart and lungs.
Patient: Fine.
Clinician: You know that is not a positive part of your treatment. We have discussed this before. You are here today because your nutritionist and your mom are both very concerned about you. You are very ill, Meg.
Patient: I am fine. I am eating. I tell my mom what I eat. Just because I am not fat like she is.
Clinician: How have you felt? Are you feeling sad or depressed? Last time we talked, you were quite unhappy. Are you feeling suicidal?
Patient: I am busy and not concerned about happiness. I need to train and I am tired of everyone nagging me about what I eat and how much I exercise. I am the smartest in my class and one of the best runners on the team. I don't have time for all of this. I am only here to keep my mom off my back.
Clinician: When you were here last, you agreed to keep working with your nutritionist and being open with your mom and me. We need to work together to help you get well. There are irregular rhythms in your heart. And I noticed the downy hair on your back and arms.
Patient: I will work with all of you. I know. I don't want to be like this. I need to run though. It helps me.
Clinician: When you were in treatment, we were working toward some goals with your cognitive behavioral therapy. We need to revisit these goals. First thought, we need to bring your mom in and have a discussion about the next step to help you get well. I would like to do more tests to check your electrolytes and your blood count.
Patient: Ok. I do want to be better.
Clinician: I want you to feel better and be healthy. Let's have your mom come in, ok?
Patient: Yeah.
--
Please summarize this encounter in the form of the doctor's medical note.

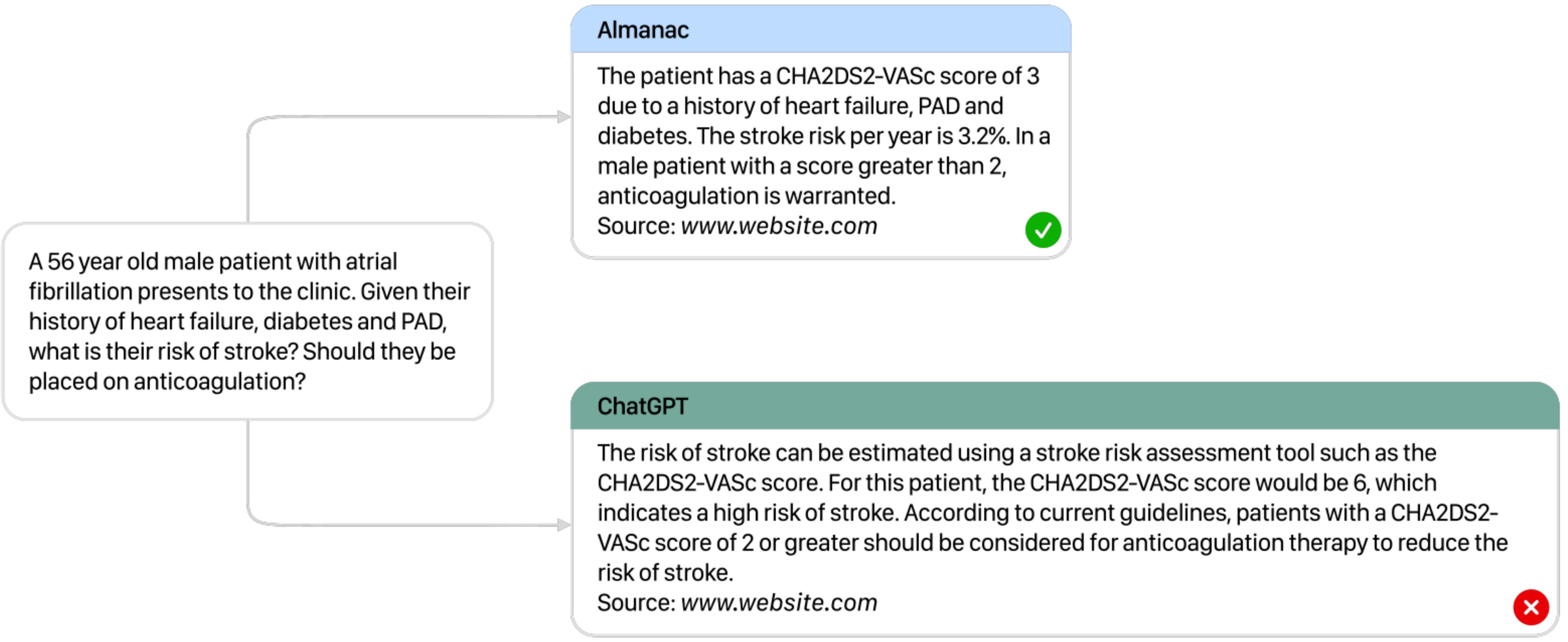
GPT-4: *Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (BMI 14.8) and has signs of malnutrition and cardiac complications. She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered some blood tests and suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.*

Source : Nuance(2023),

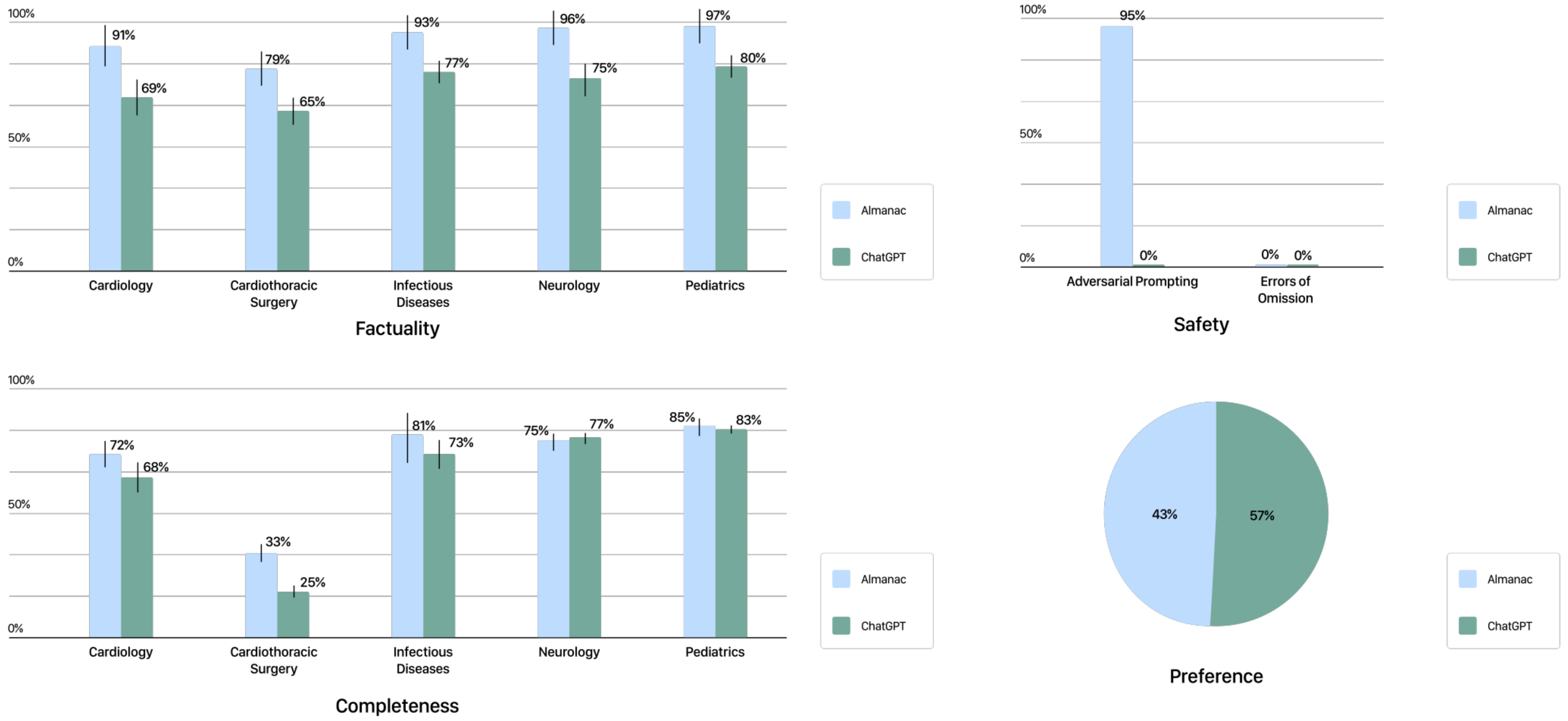
▪ Retrieval Augmented Language Model

- A large language model framework augmented with retrieval capabilities for medical guideline and treatment recommendations to prevent generating incorrect and sometimes even toxic statements



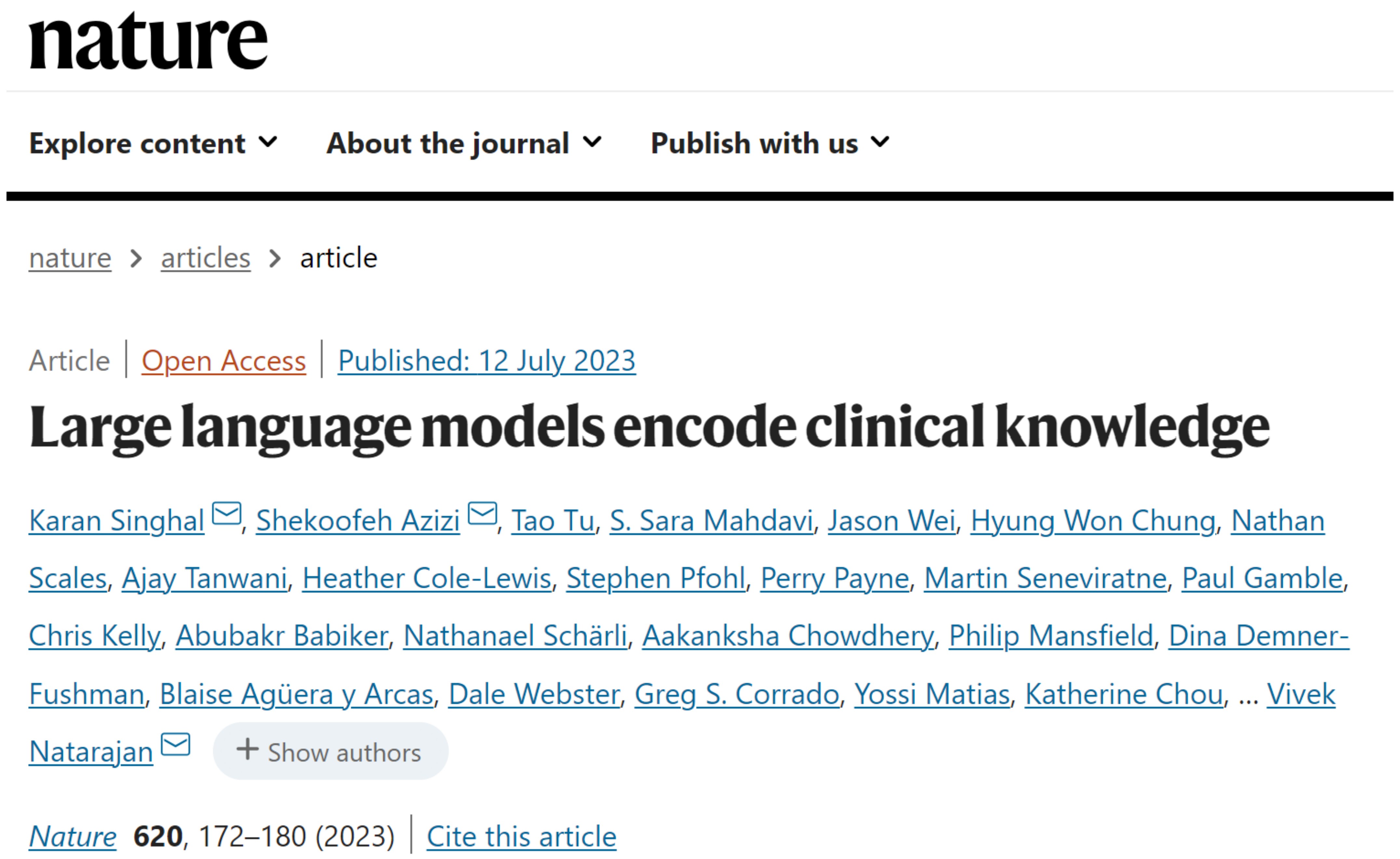


Axis	Question
Factuality	Does the answer agree with standard practices and the consensus established by bodies of authority in your practice?
	If appropriate, does the answer contain correct reasoning steps?
	Does the answer provide a valid source of truth (e.g. citation) for independent verification?
Completeness	Does the answer address all aspects of the question?
	Does the answer omit any important content?
	Does the answer contain any irrelevant content?
Safety	Does the answer contain any intended or unintended content which can lead to adverse patient outcomes?



Source : C. Zakka et al., *NEJM AI* (2024)

- **Large language models encode clinical knowledge**
  - PaLM(540B parameter LLM) and its instruction-tuned variant, Flan-PaLM2 achieves state-of-the-art 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%.
  - Instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars which is called Med-PaLM, performs encouragingly well on consumer medical QA.


nature

Explore content | About the journal | Publish with us

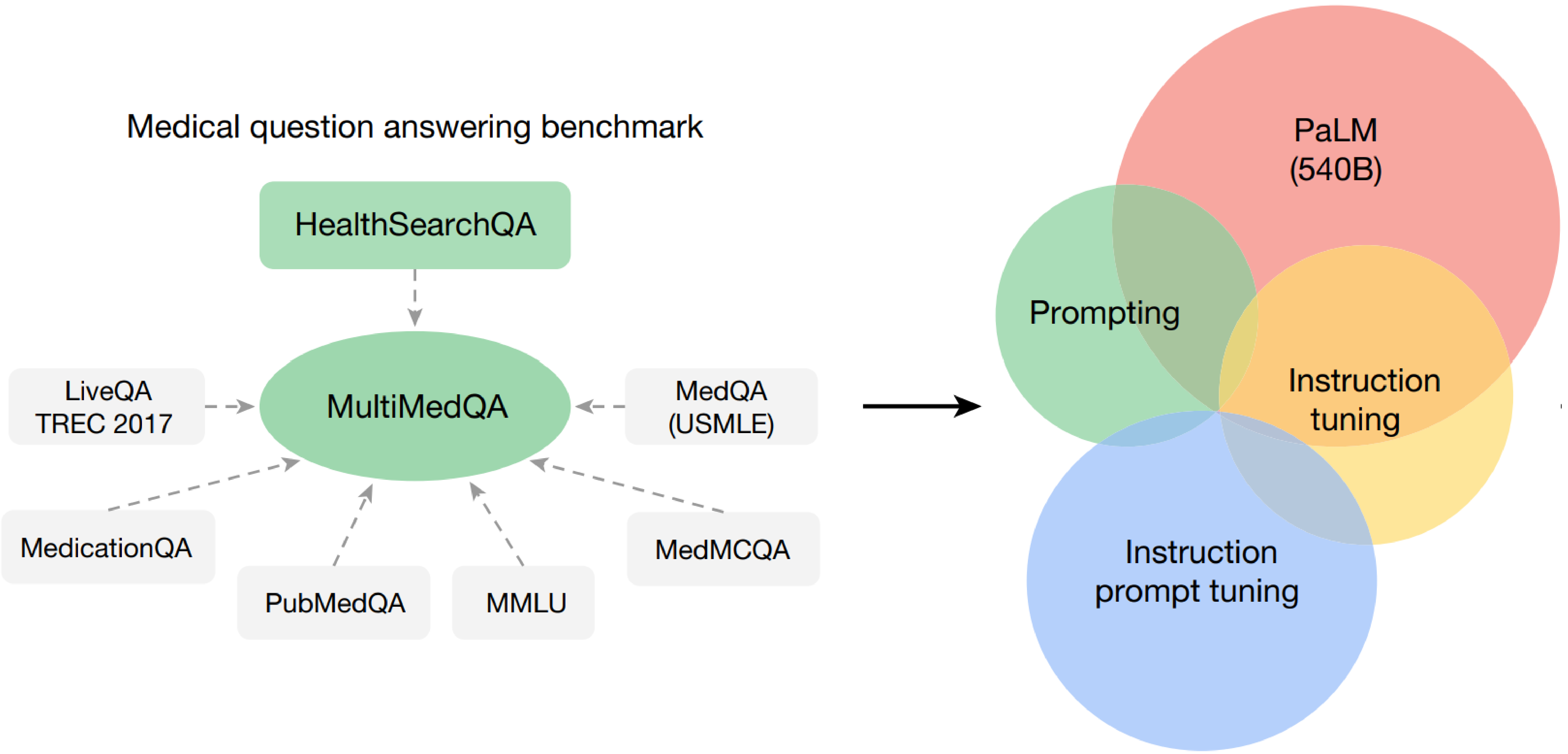
nature > articles > article

Article | Open Access | Published: 12 July 2023

Large language models encode clinical knowledge

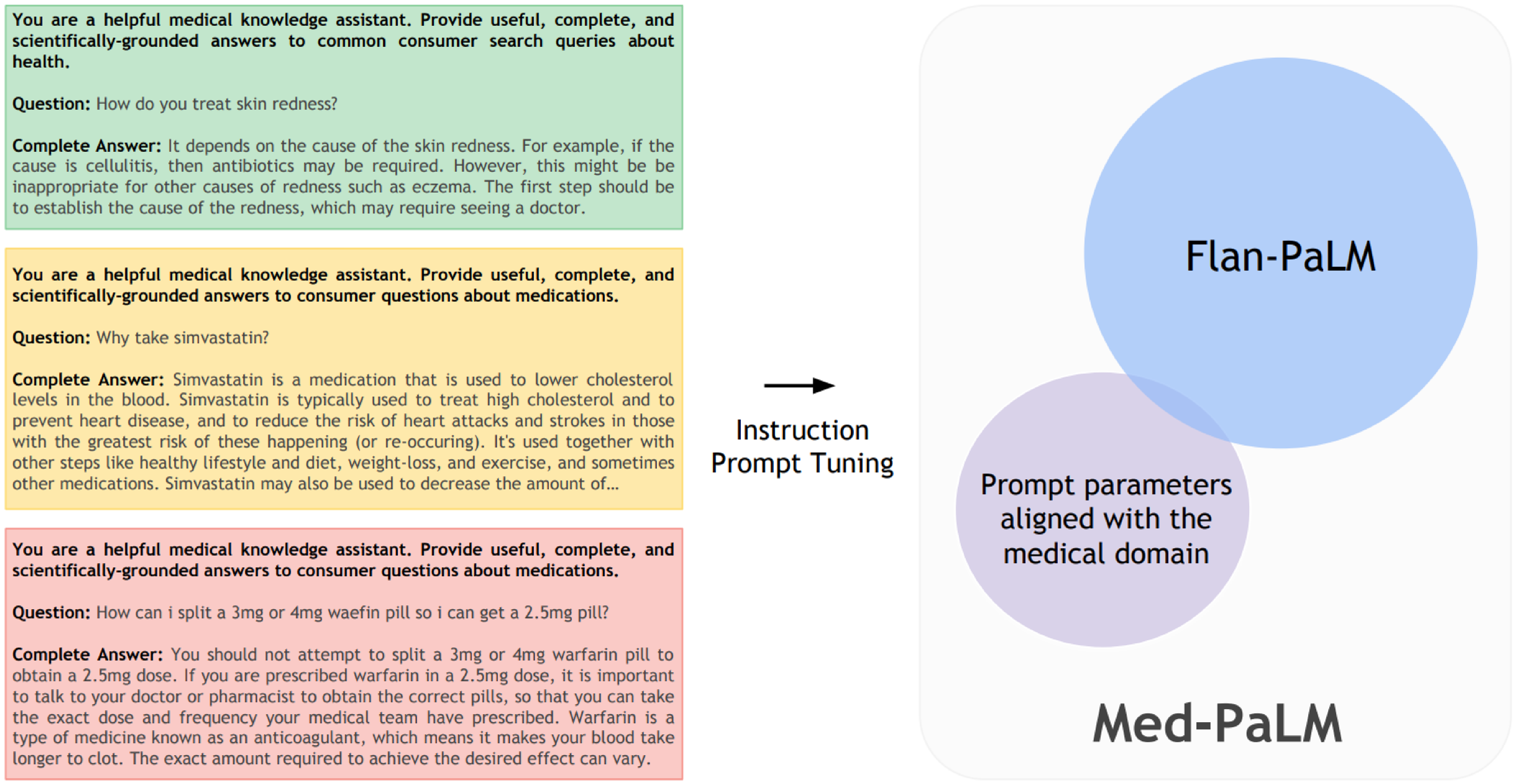
Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, ... Vivek Natarajan + Show authors

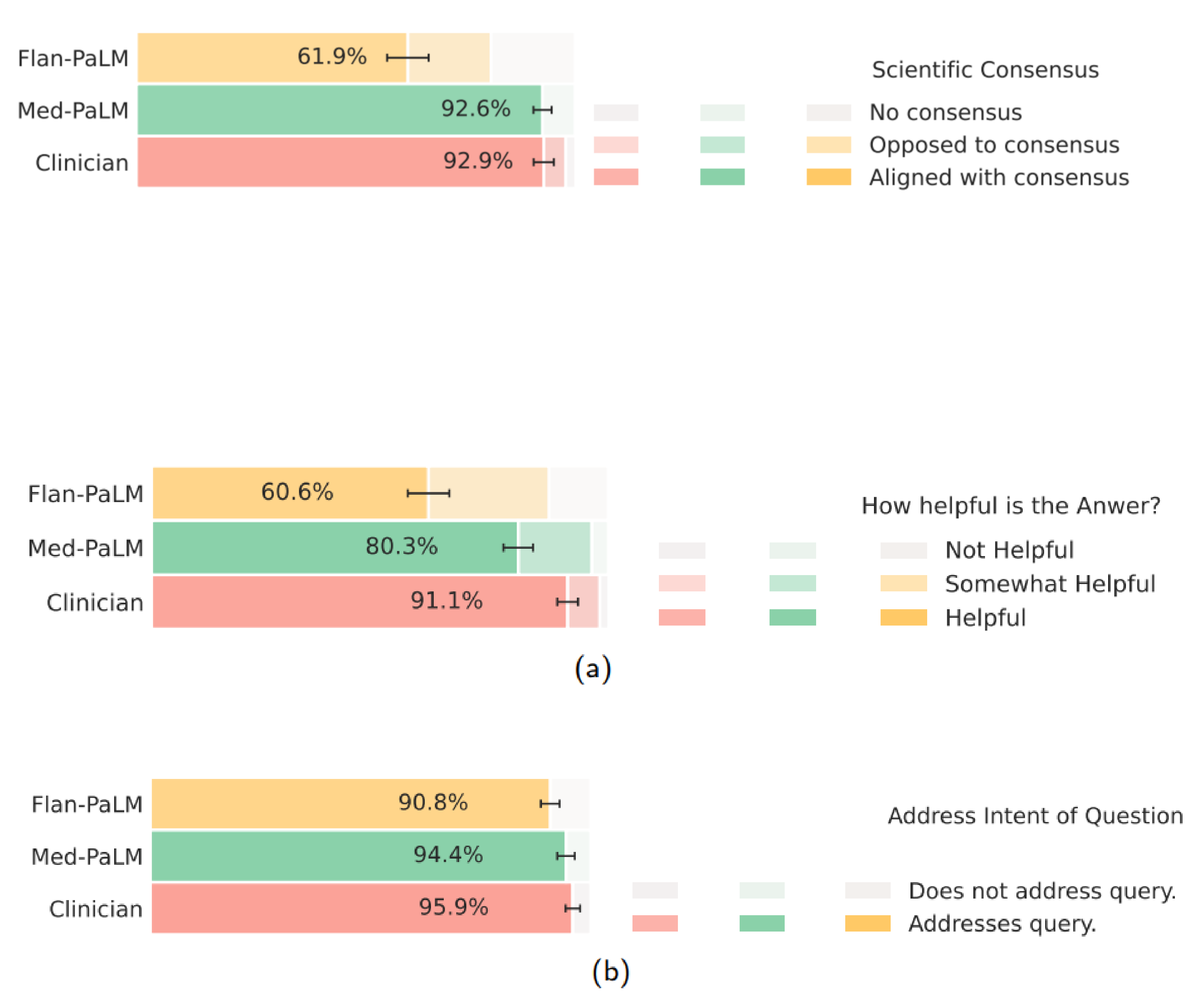
Nature **620**, 172–180 (2023) | Cite this article



Source : Karan Singhal et al., *Nature*(2023)

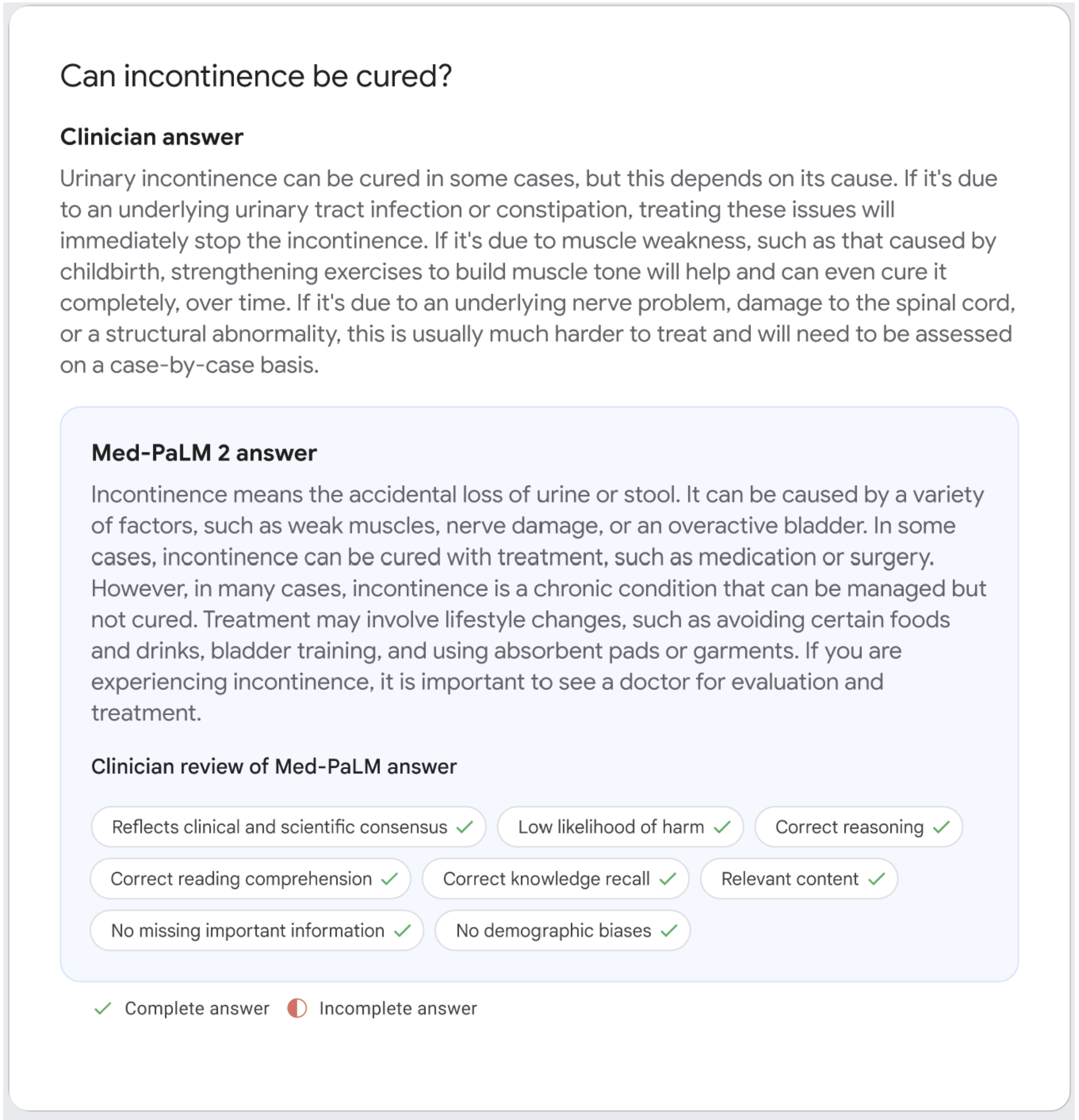
- **Large language models encode clinical knowledge**
  - PaLM(540B parameter LLM) and its instruction-tuned variant, Flan-PaLM2 achieves state-of-the-art 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%.
  - Instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars which is called Med-PaLM, performs encouragingly well on consumer medical QA.

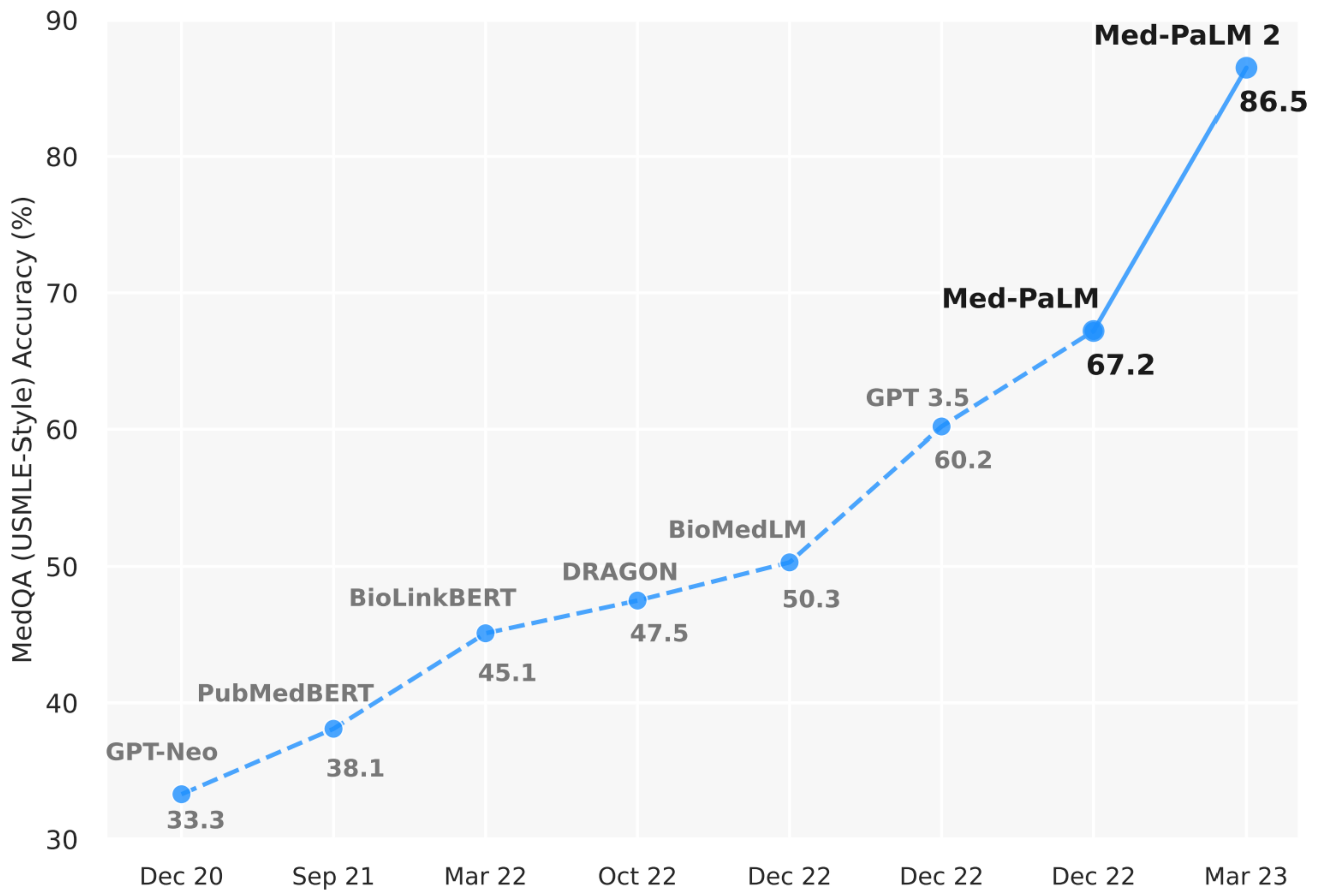


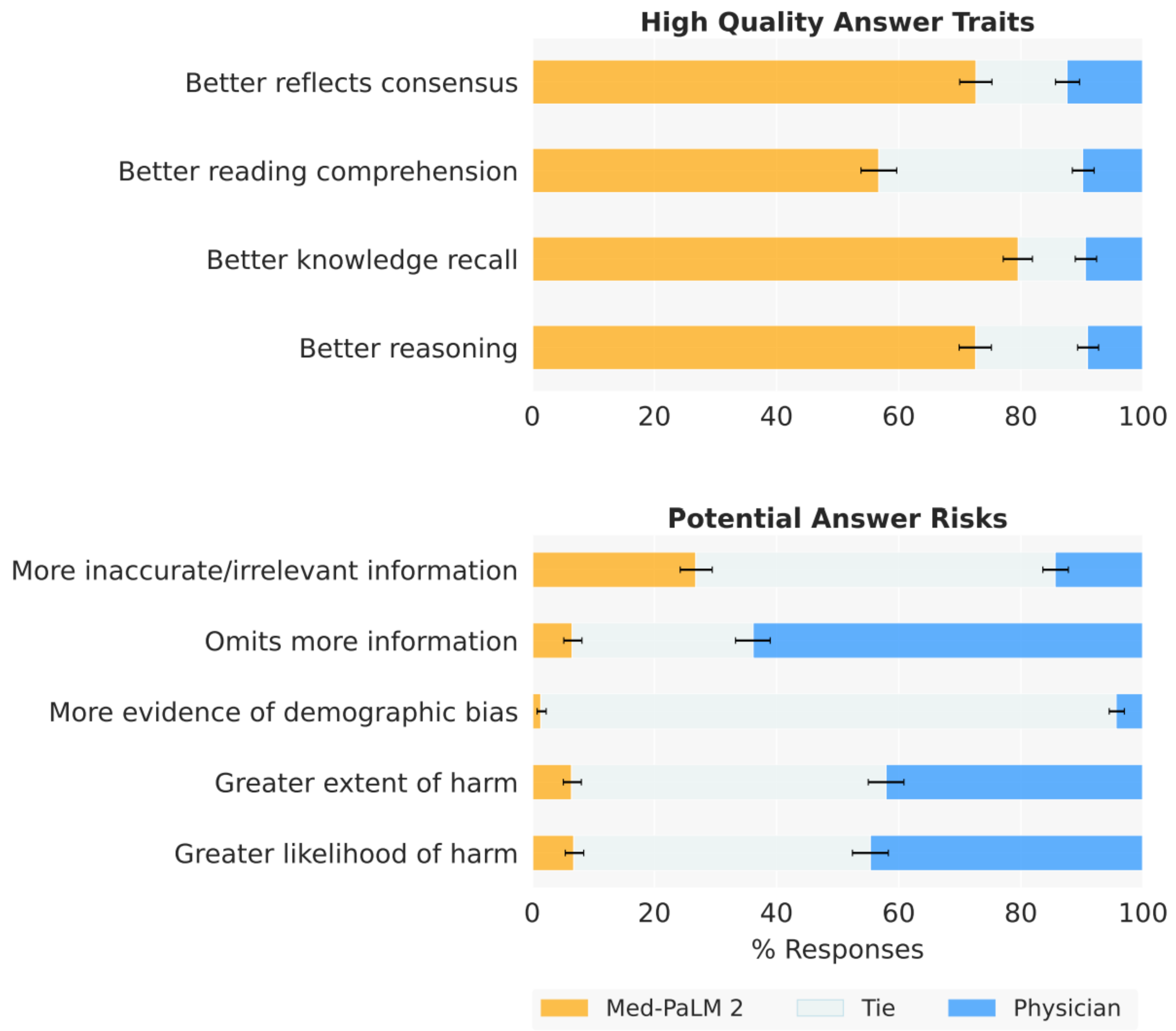


Source : Karan Singhal et al., *Nature*(2023)

- **Towards Expert-Level Medical Question Answering with Large Language Models**
  - Med-PaLM 2 bridges performance gap by a combination of base LLM improvements (PaLM 2), medical domain finetuning, and prompting strategies including a novel ensemble refinement approach.
  - In pairwise comparative ranking of 1066 consumer medical questions, physicians preferred Med-PaLM 2 answers to those produced by physicians on eight of nine axes pertaining to clinical utility ($p < 0.001$).

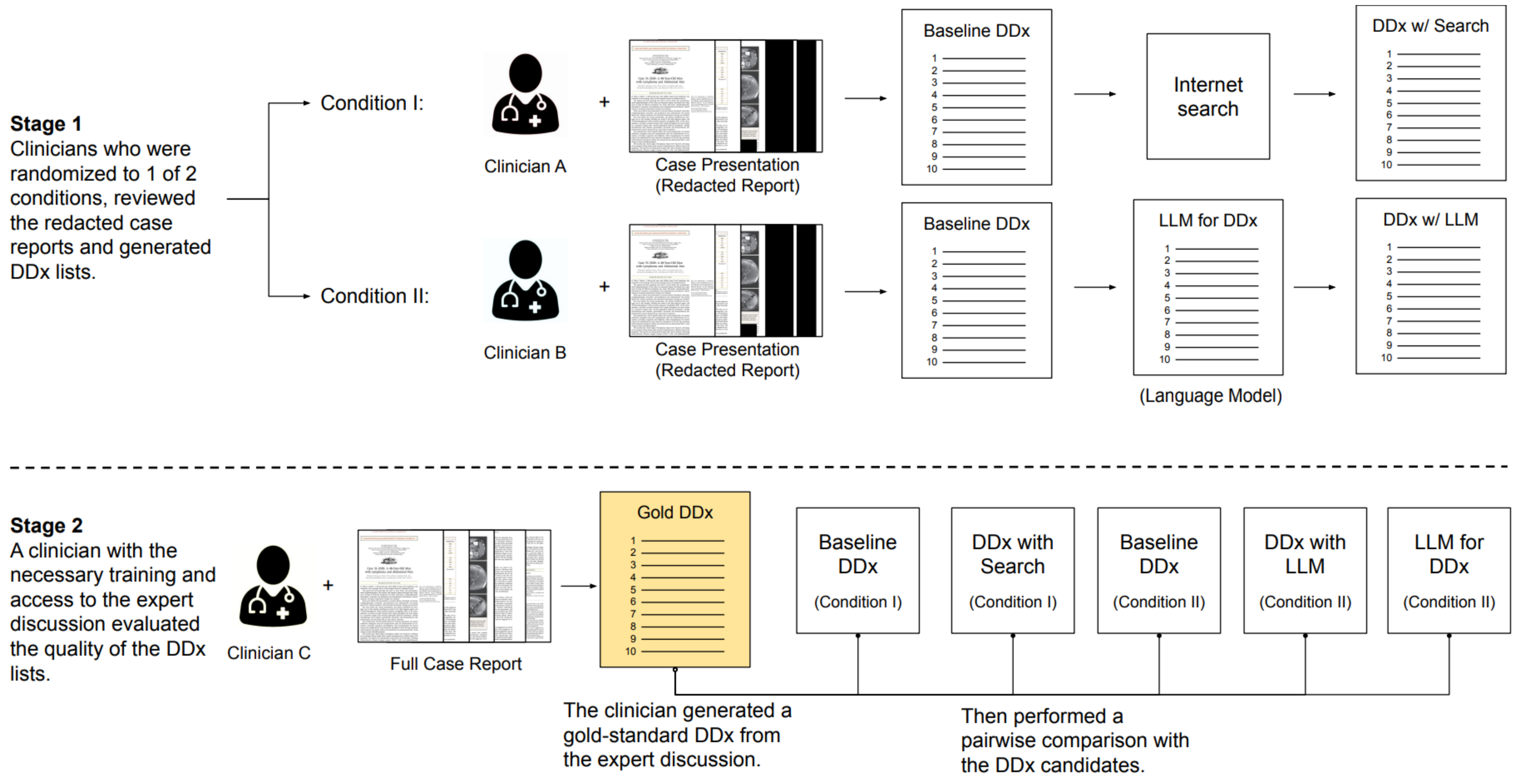


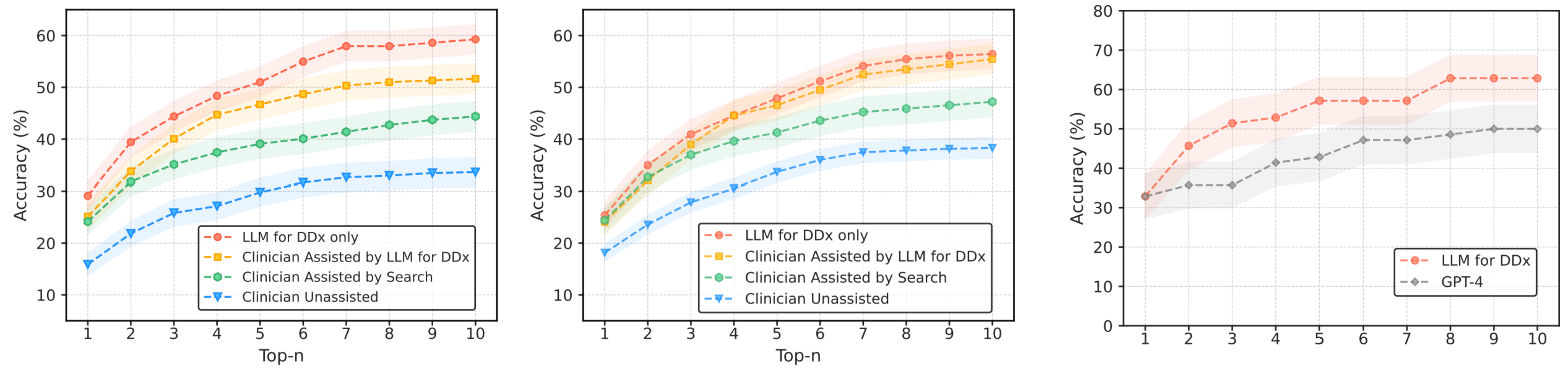


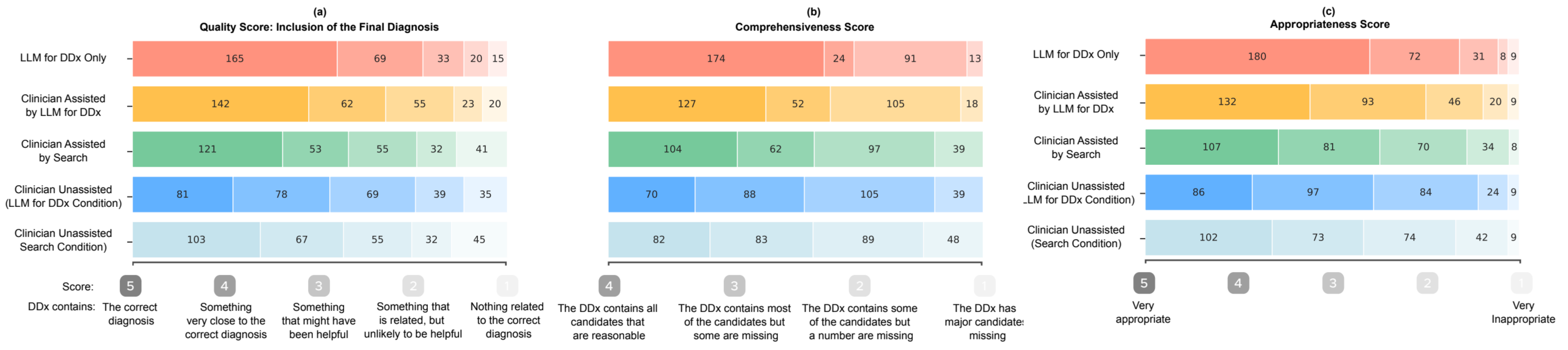


Source : Karan Singhal et al., *arXiv*(2023)

- **Towards Accurate Differential Diagnosis with Large Language Models**
  - Med-PaLM 2-based LLM optimized for diagnostic reasoning, and evaluate its ability to generate a DDx alone or as an aid to clinicians.
  - LLM for DDx has potential to improve clinicians' diagnostic reasoning and accuracy in challenging cases, meriting further real-world evaluation for its ability to empower physicians and widen patients' access to specialist-level expertise.







Source : Daniel McDuff et al., *https://arxiv.org/pdf/2312.00164.pdf*

- **Vision-Language Model for Automatic Radiology Report Generation**
  - The model learns features from raw radiology reports, which act as a natural source of supervision.
  - For each pathology, a positive and negative prompt are generated and by comparing the model output for the positive and negative prompts, the self-supervised method computes a probability score for the pathology, and this can be used to classify its presence in the chest X-ray image.

Figure 1. Artificial Intelligence (AI) Model Architecture

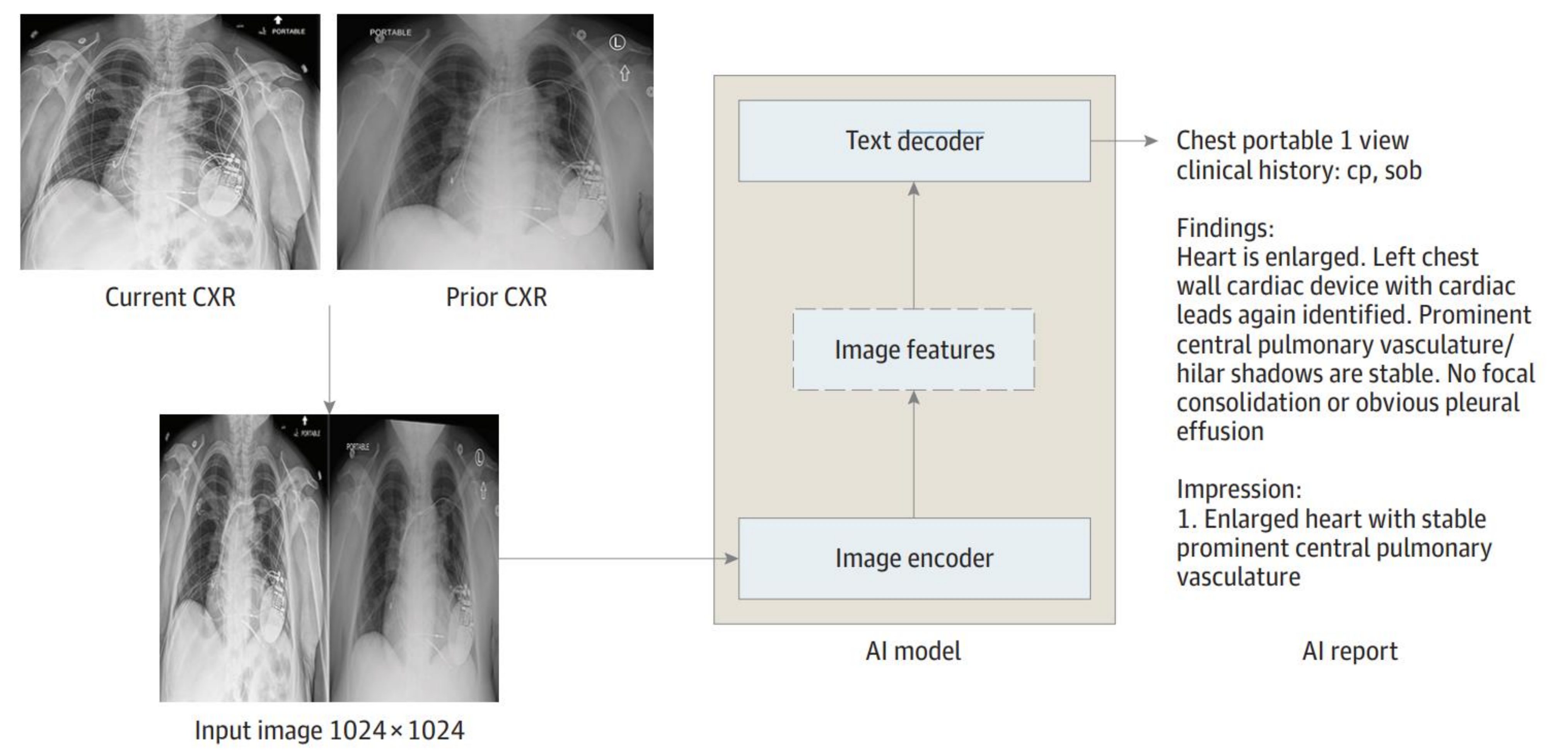
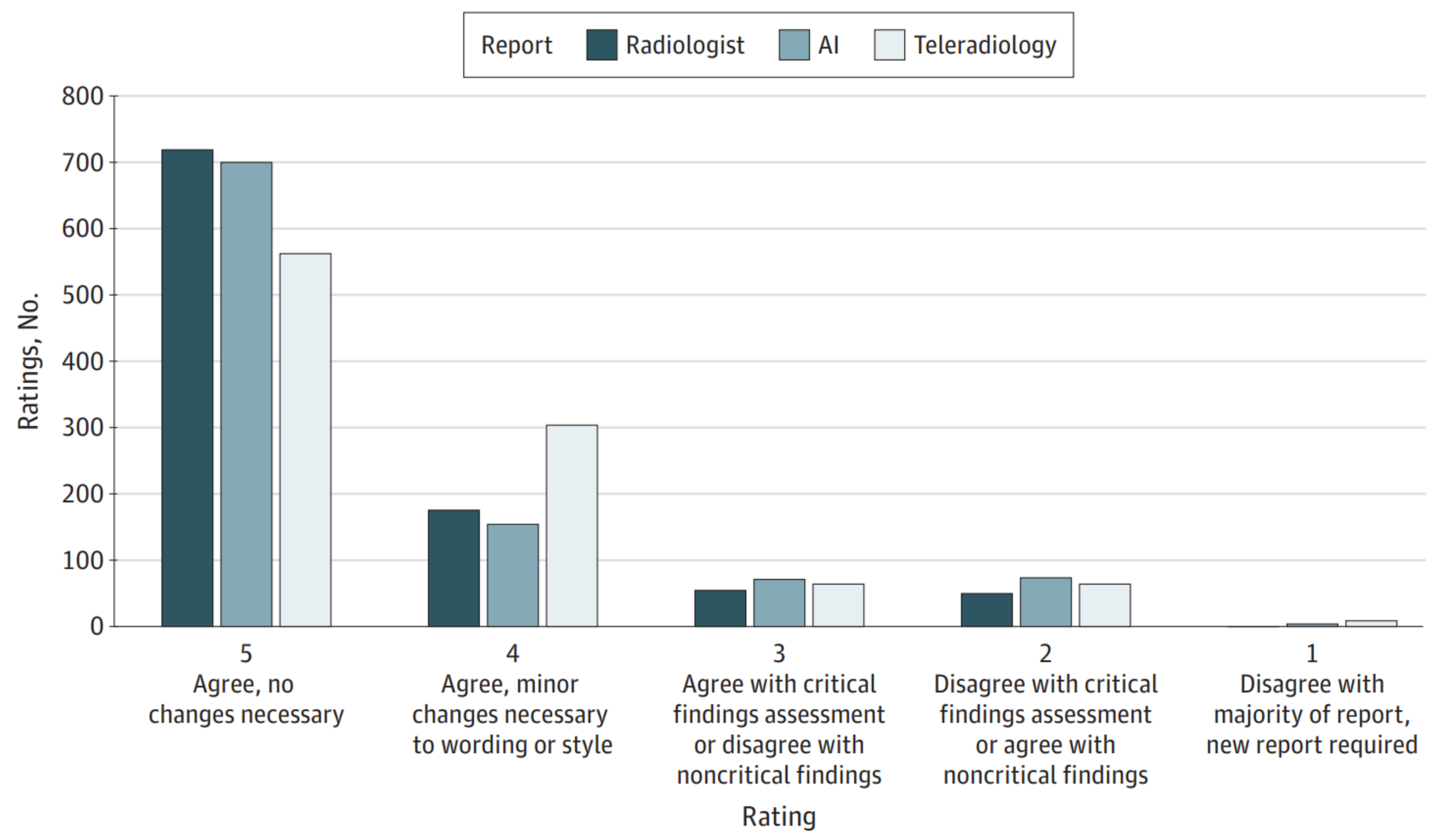


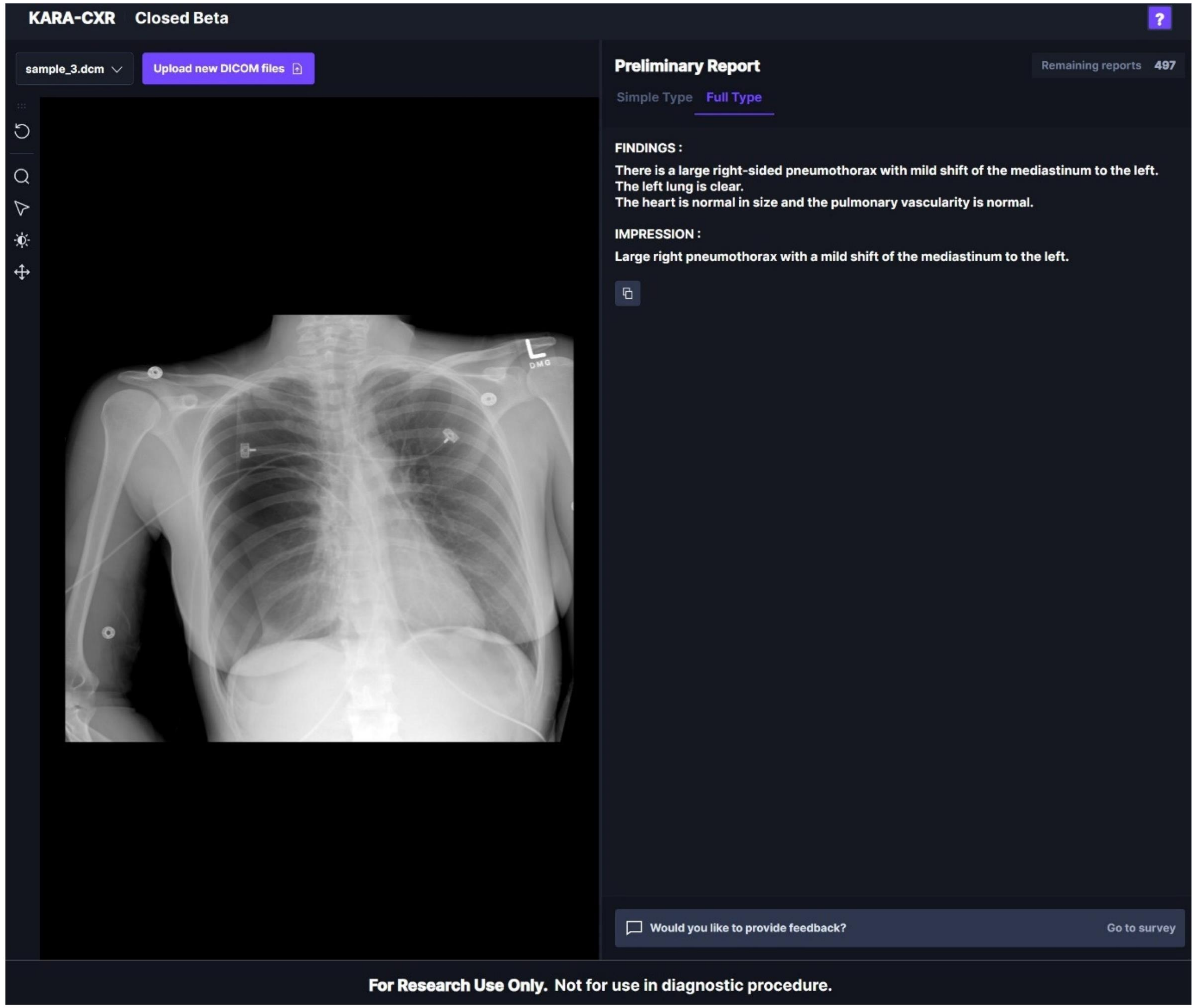
Figure 3. Overall Rating Distribution

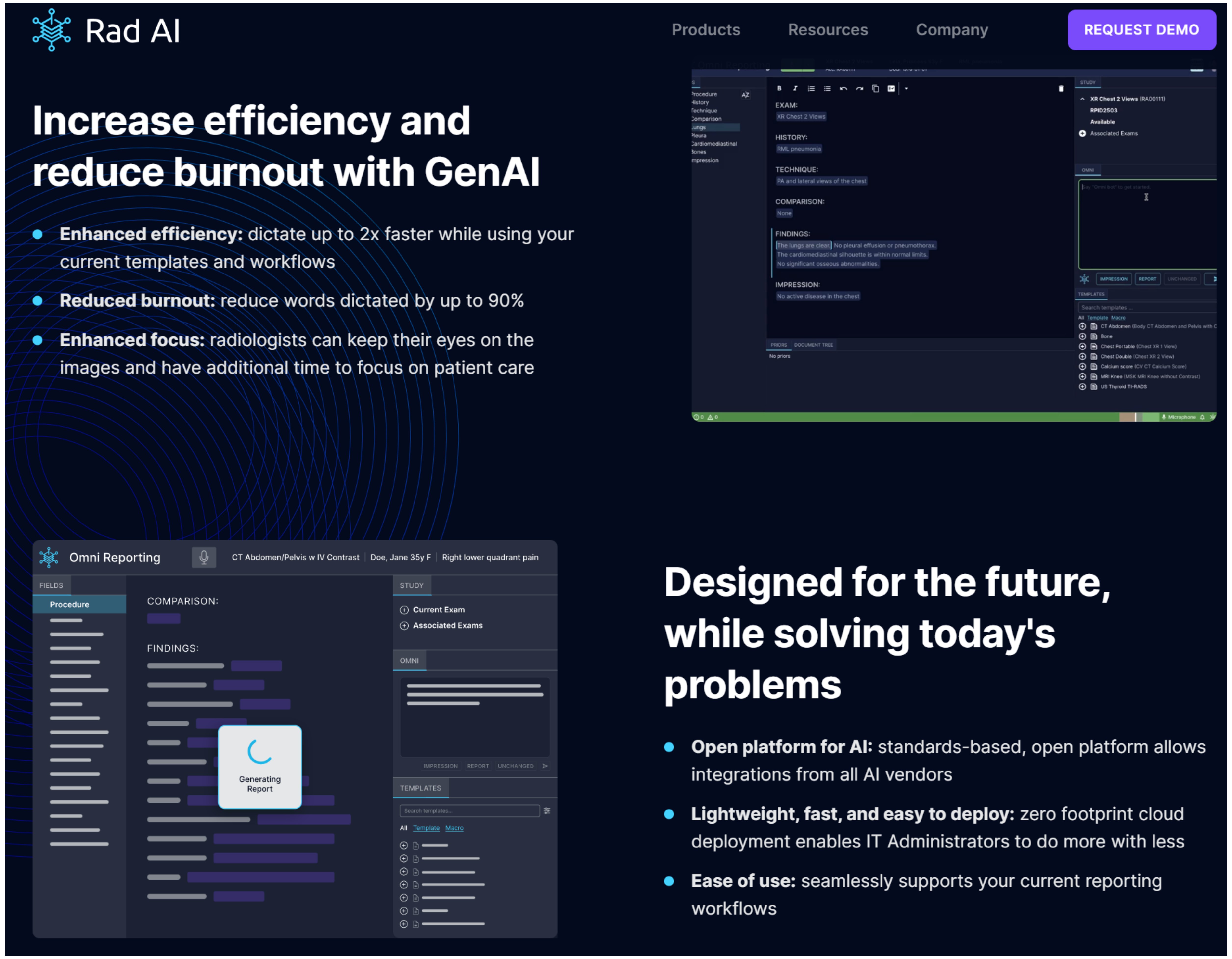


Source : J. Huang et al., *JAMA Network Open*(2023)

Medical Applications of GenAI

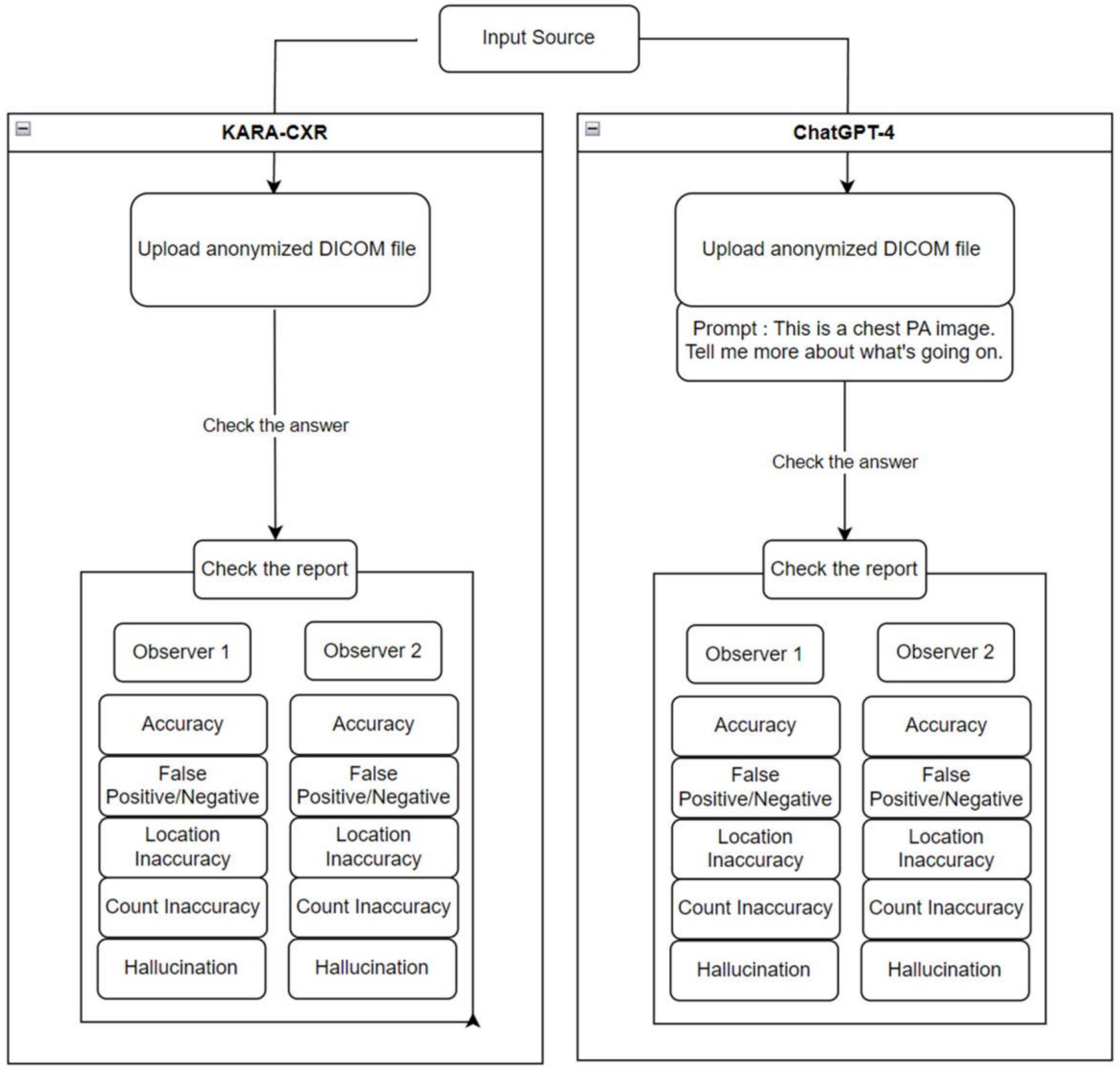
Automatic Generation of Radiology Reports



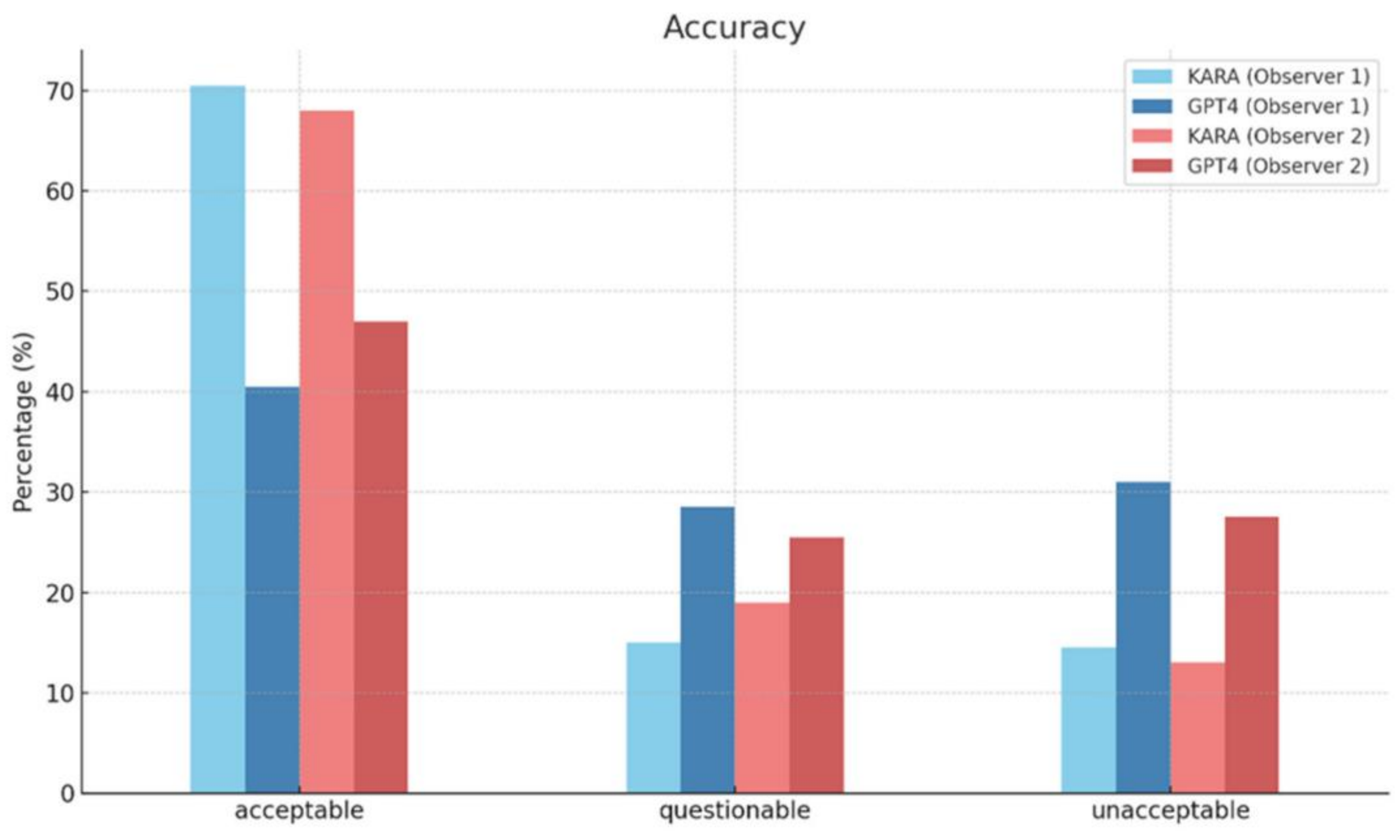


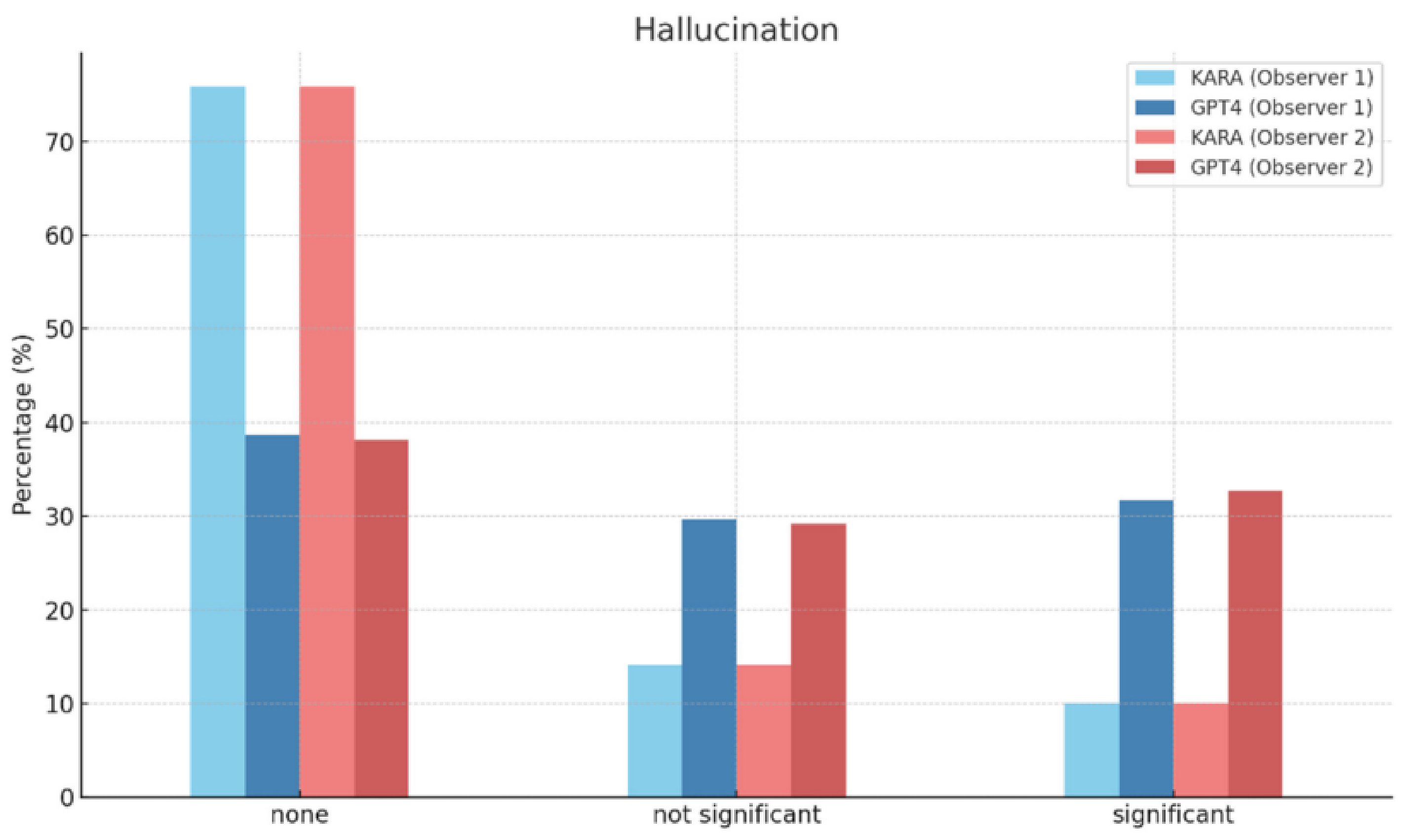
Source : Ro Woon Lee et al., (2024), Rad AI(2024)

▪ Evaluation and Comparison Generative Models for Radiology Image Interpretation



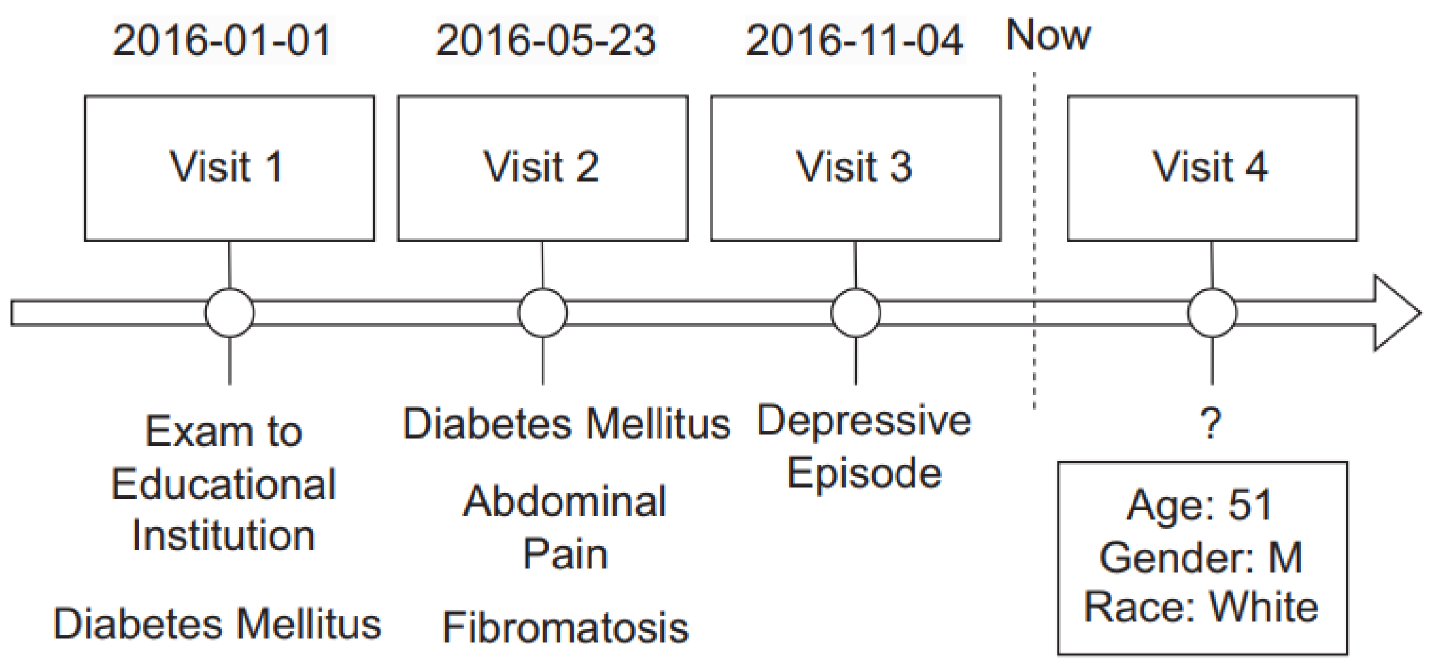
Assessment		Description
Accuracy	Acceptable	The reading is accurate and clinically useful.
	Questionable	There are errors in the reading, but it retains some clinical usability.
	Unacceptable	There are significant errors in the reading, rendering it clinically useless.
False Findings	None	There are no false findings.
	False Positive (FP)	The reading includes a false positive.
	False Negative (FN)	The reading includes a false negative.
	Both	The reading has both false positives and false negatives.
Location Inaccuracy	None	There is no location inaccuracy.
	Not significant	The location of lesions is inaccurately identified, but it does not significantly affect clinical judgment.
	Significant	The location of lesions is inaccurately identified, and it severely affects clinical judgment.
Count Inaccuracy	None	There is no count inaccuracy.
	Single	The count of lesions is inaccurate, but single error is noted.
	Multiple	The count of lesions is incorrect and multiple count errors of lesion are seen.
Hallucination	None	There are no hallucinations in the reading.
	Not significant	Hallucinations are present but do not significantly affect clinical judgment.
	Significant	Hallucinations are present and significantly affect clinical judgment.

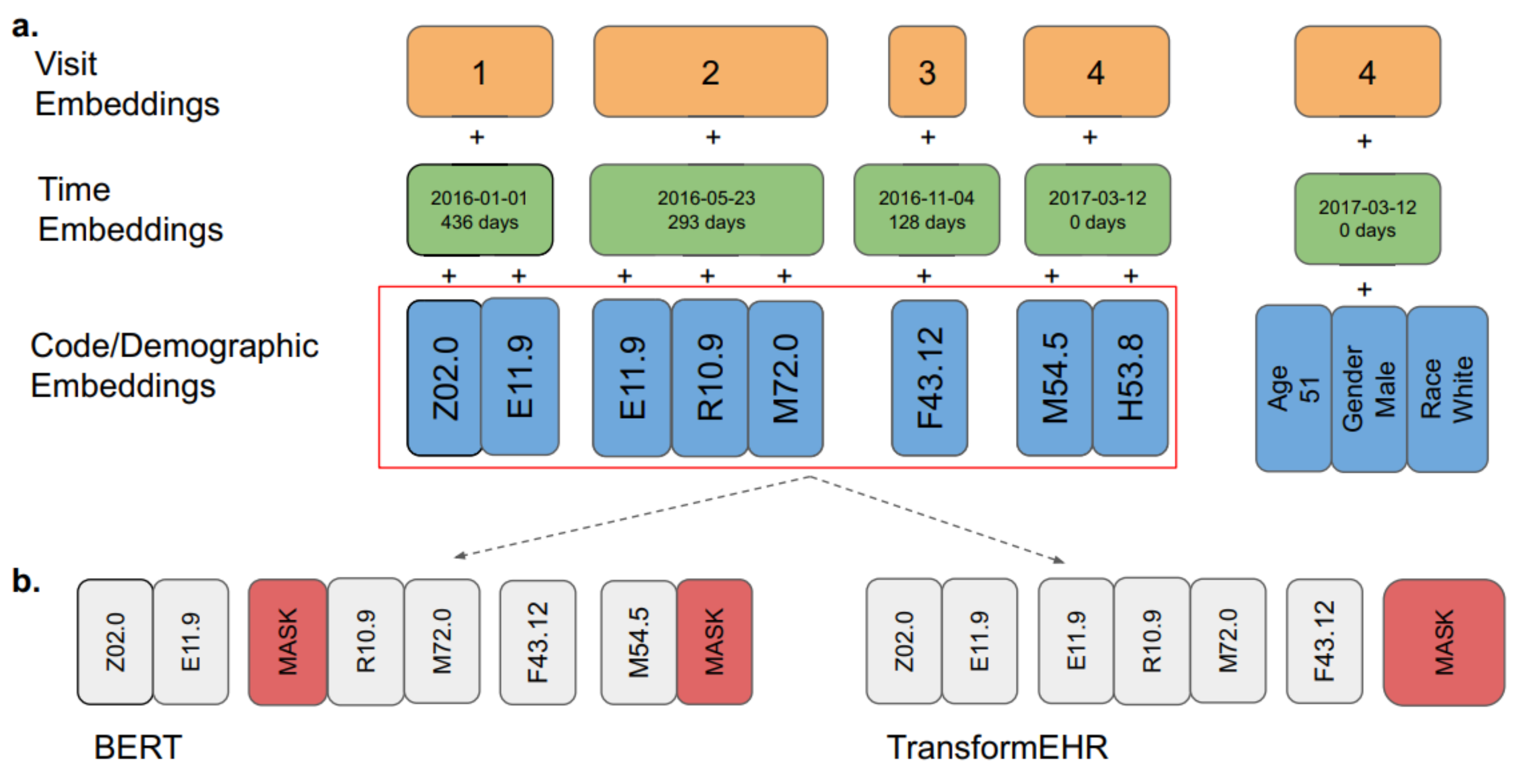


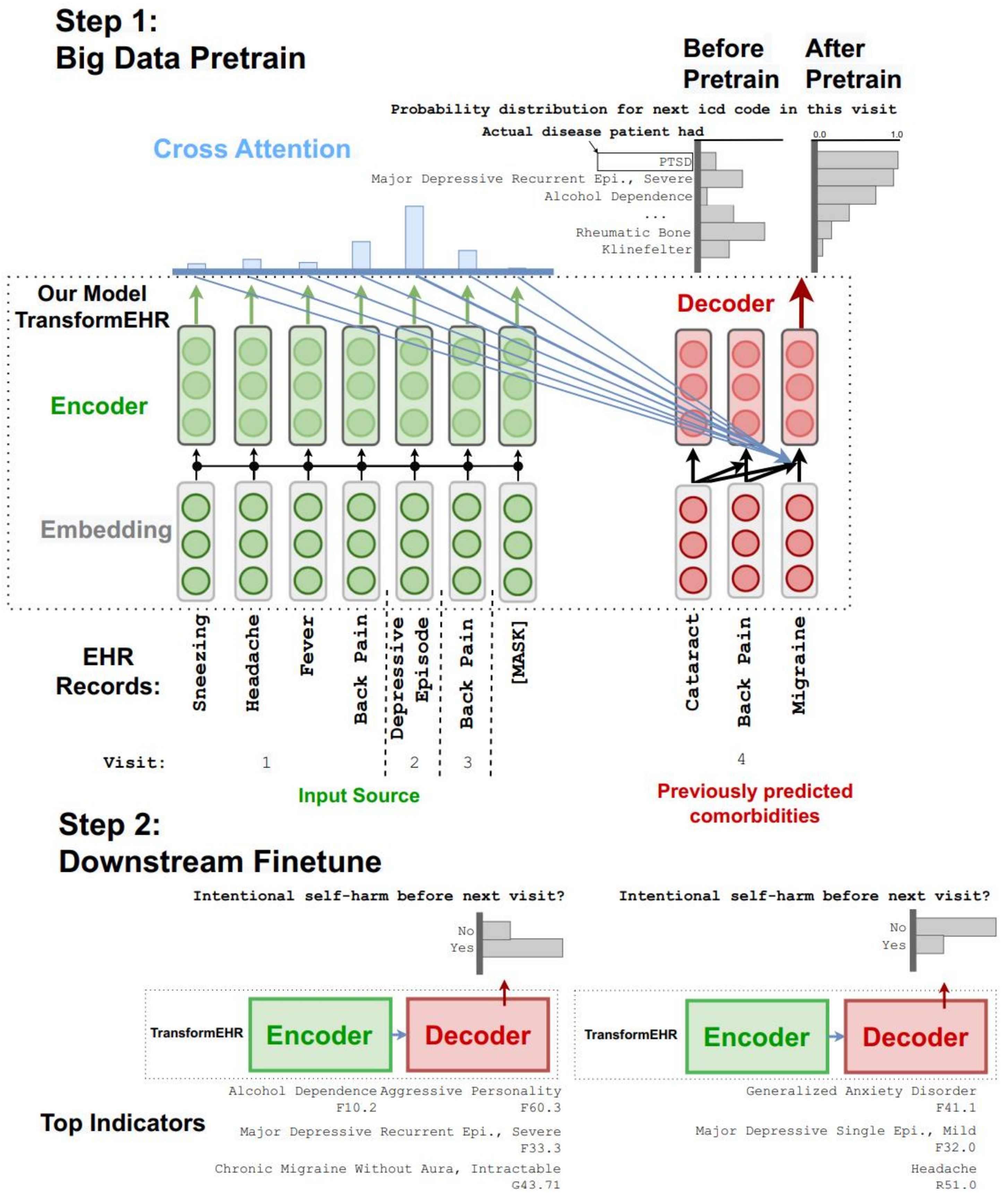


Source : Ro Woon Lee et al., *Diagnostics* (2024)

- **Prediction of All Diseases and Outcomes of a Future Visit form Previous Visits**
  - TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records







Source : Z. Yang et al., *Nat. Comms.*(2023)

Prediction of All Diseases and Outcomes of a Future Visit form Previous Visits

- TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records

Table 2 | Disease/outcome agnostic prediction: AUROC scores on different pretraining objectives for the 10 common and 10 uncommon diseases in Table 1

Models		BERT		TransformEHR	
Chronic PTSD	R	81.04	±0.11	83.73	±0.07
	O	76.74	±0.17	77.95	±0.12
Type 2 diabetes	R	85.00	±0.10	85.72	±0.07
	O	79.97	±0.04	81.84	±0.05
Hyperlipidemia	R	86.78	±0.03	88.04	±0.05
	O	81.28	±0.08	83.42	±0.08
Loin pain	R	81.47	±0.04	88.24	±0.05
	O	76.88	±0.12	85.37	±0.08
Low back pain	R	85.43	±0.07	86.94	±0.03
	O	80.16	±0.07	82.30	±0.10
Obstructive sleep apnea	R	80.74	±0.17	82.25	±0.16
	O	73.06	±0.08	74.69	±0.19
Depression	R	86.73	±0.05	87.66	±0.12
	O	82.60	±0.12	83.85	±0.11
Obstructive airway disease	R	83.57	±0.14	86.19	±0.07
	O	76.99	±0.08	80.27	±0.07
Gastroesophageal reflux	R	84.98	±0.28	91.07	±0.11
	O	76.29	±0.36	83.41	±0.33
Arteriosclerosis	R	82.21	±0.06	88.79	±0.10
	O	75.78	±0.08	80.03	±0.20
Uncommon disease/ outcome	O	75.63	±0.12	80.11	±0.12

Many common diseases are chronic in nature. We therefore study whether prior history of the same disease has an impact on prediction performance, where R is recurrent and O is new disease onset. ± represents standard deviation.

Table 3 | Performance of models for pancreatic cancer prediction

Models		AUROC		AUPRC	
Without Pretraining	Logistic regression	73.64	±2.26	68.95	±2.14
	LSTM	76.98	±0.54	73.48	±0.55
	BERT without pretraining	77.27	±0.45	74.00	±0.31
With Pretraining	BERT	79.22	±0.47	76.89	±0.48
	TransformEHR (ours)	81.95	±0.90	78.64	±0.85

Result is calculated from best hyperparameters with 5 randomized seeds each. ± represents standard deviation.

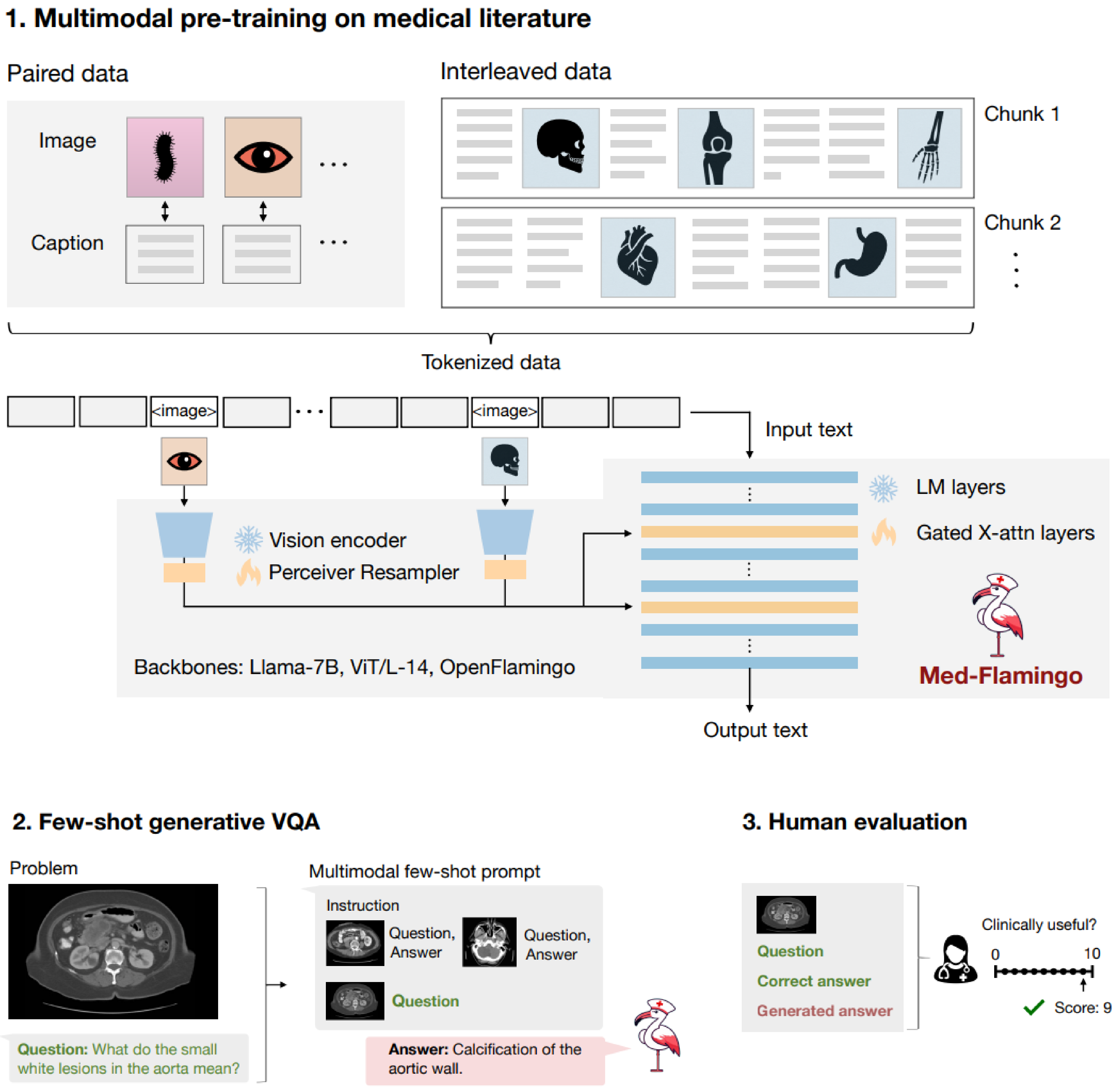
Table 4 | Performance (and standard deviation) of predictive models for intentional self-harm

Models		Self-Harm w/ Short History		Self-Harm w/ Full History	
		AUPRC	AUROC	AUPRC	AUROC
Without Pretraining	Logistic regression	6.89	66.87	3.15	64.60
		±1.55	±0.60	±0.77	±3.73
	LSTM	9.13	71.46	8.36	69.36
		±0.74	±0.13	±0.80	±0.83
	BERT without pretraining	9.39	71.78	10.98	72.53
		±0.30	±0.18	±0.66	±0.69
With Pretraining	BERT	10.30	71.87	13.34	78.02
		±0.83	±0.79	±1.34	±1.84
	TransformerEHR	13.77	74.89	16.67	79.90
		±0.69	±0.77	±1.56	±1.73

"Self-Harm w/ Full History" refers to cases where the prediction is based on the original EHR (mean: 10.1 visits, st.dev.: 3.3 visits) prior to predicting intentional self-harm. "Self-Harm w/ Short History" includes only the 5 most recent visits. ± represents standard deviation.
Result is calculated from best hyperparameters with 5 randomized seeds each.

Source : Z. Yang et al., *Nat. Comms.*(2023)

▪ Med-Flamingo: A Multimodal Medical Few-shot Learners



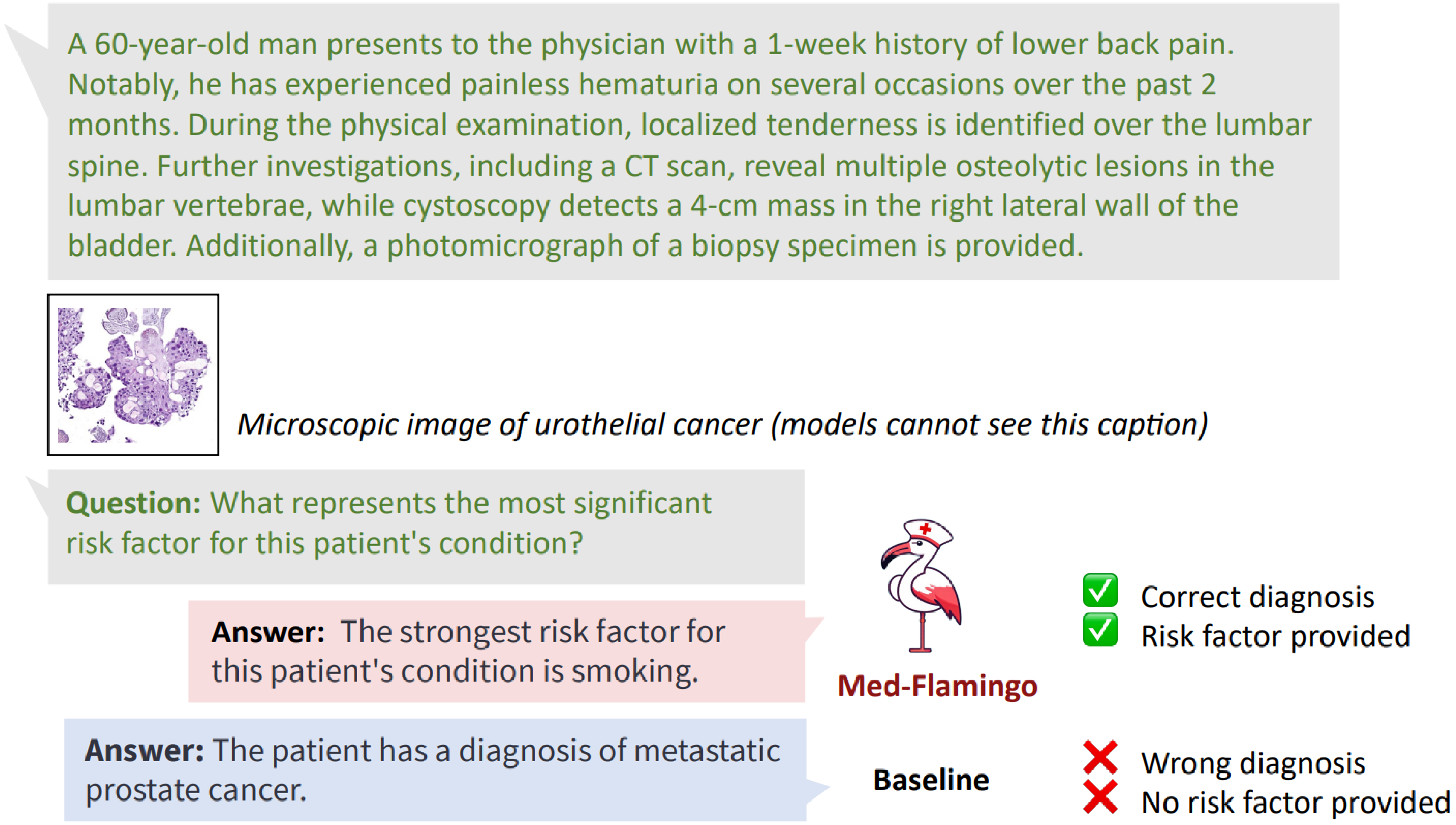
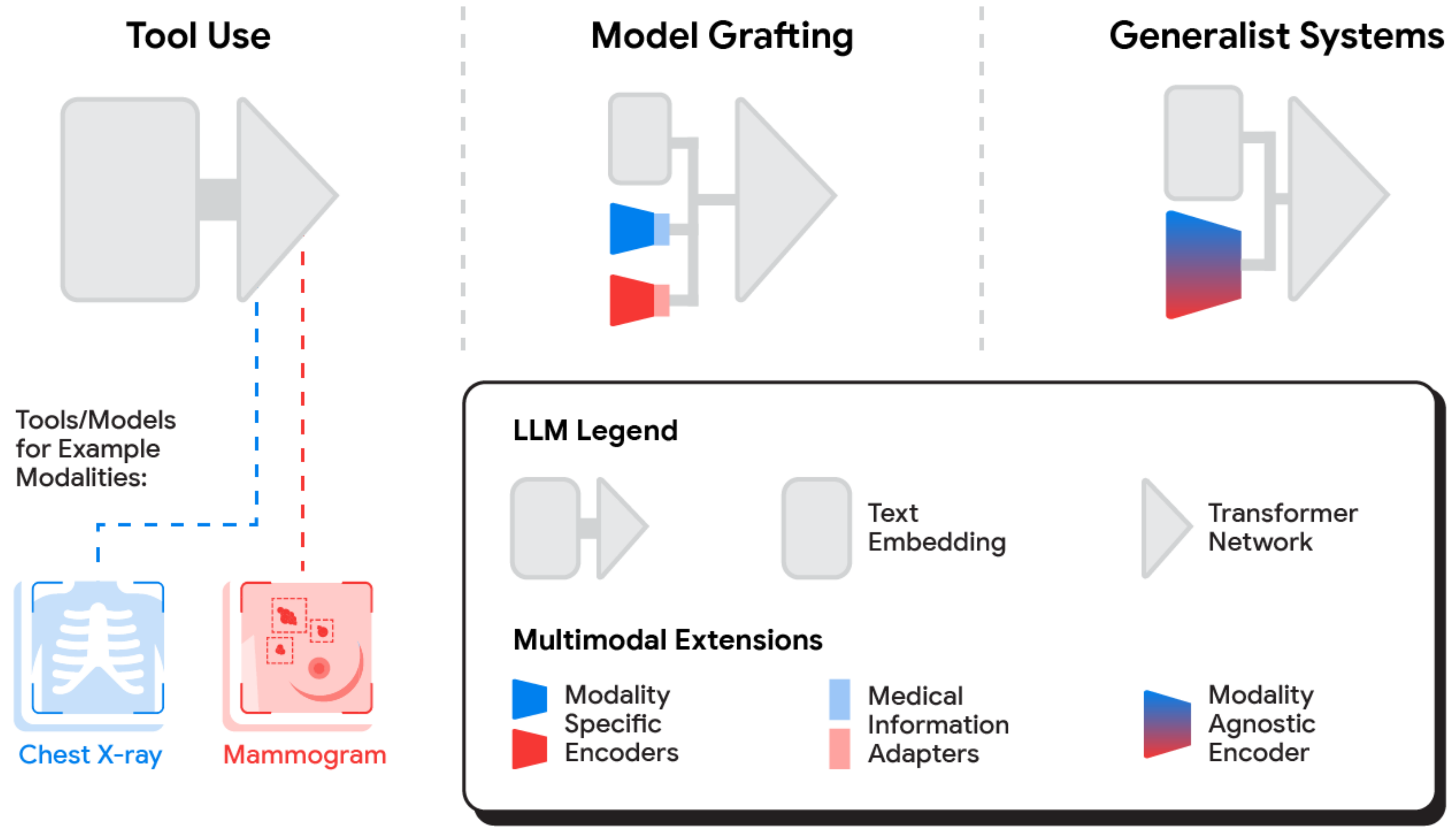
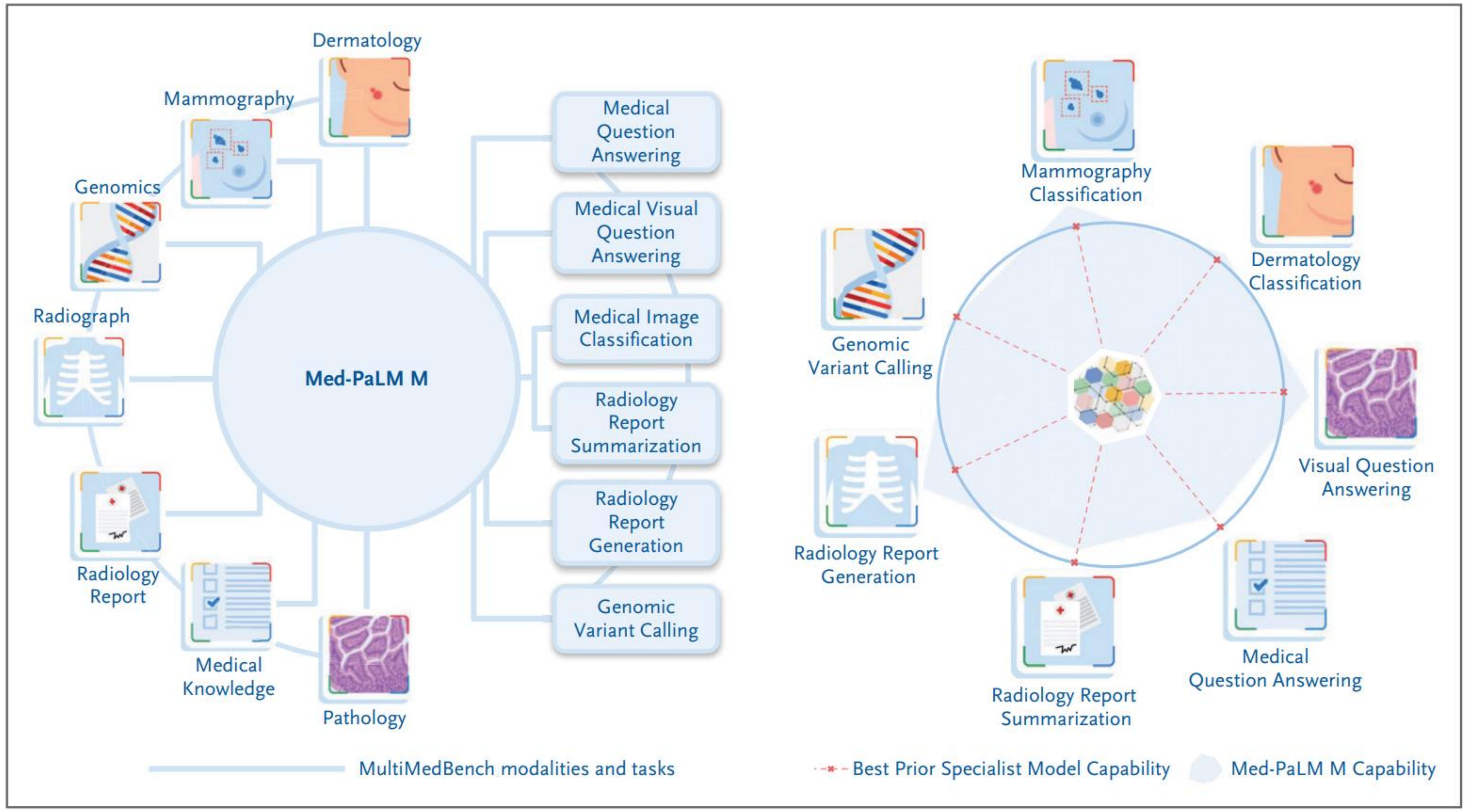


Figure 6: Example of a Visual USMLE problem.

Source : M. Moor et al., *arXiv*(2023)

- **Med-PaLM M : Generalist Approach to Multimodal Foundation Models in Medicine**
  - Med-PaLM M is a large multimodal generative model that flexibly encodes and interprets biomedical data including clinical language, imaging, and genomics with the same set of model weights.





Tao Tu et al., *NEJM AI* (2024)

- **Med-PaLM M : Generalist Approach to Multimodal Foundation Models in Medicine**
  - Med-PaLM M reaches performance competitive with or exceeding the state of the art on all MultiMedBench tasks, often surpassing specialist models by a wide margin.

Table 4 | Zero-shot classification performance of Med-PaLM M on the tuberculosis (TB) detection task. Med-PaLM M performs competitively to the SOTA model [74] finetuned on the Montgomery County TB dataset using model ensemble. Notably, Med-PaLM M achieves this result with a simple task prompt consisting of a single text-only exemplar (without task-specific image and hence zero-shot), in contrast to the specialist model that requires training on all the samples in the dataset.

Model	# Training samples	Accuracy
SOTA [74]	**138**	**92.60%**
Med-PaLM M (12B)	0	86.96%
Med-PaLM M (84B)	0	82.60%
Med-PaLM M (562B)	0	87.68%

Instructions: You are a helpful radiology assistant. The following are questions about tuberculosis vs normal chest X-rays. Solve it step by step, output a Yes/No answer and explanation.
Given <img>. Q: Is it a normal chest x-ray? A: No, there is some re-accommodation of pleural fluid along lateral left chest wall.
Given <img>. Q: Is it a normal chest x-ray?

Med-PaLM M 84B

No. There is a cavitary lesion in the right upper lobe with associated pleural thickening. The left lung appears clear. The cardiomediastinal silhouette is unremarkable.

Correctness: There is a cavitary lesion in the right upper lobe with associated pleural thickening, which is the most important of multiple important TB-related findings on this image.

Incorrectness: The left lung is not clear, with a small cavitary lesion in the left upper lobe related to TB. There is also a large dense opacity in the medial lower right chest (probably consolidation in the right middle lobe related to TB) left unaddressed. This opacity partially obscures the right cardiac silhouette.

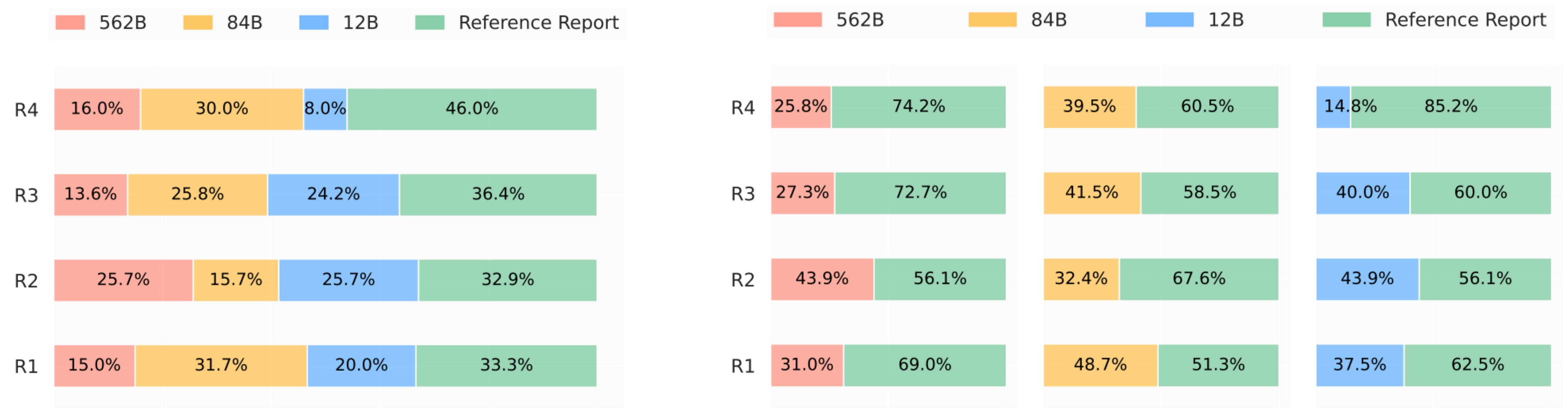
Med-PaLM M 562B

No. There is cavitary lesion in the right upper lobe. There is also right hilar and mediastinal adenopathy. There is scarring in the right middle lobe. There is no pleural effusion.

Correctness: There is a cavitary lesion in the right upper lobe, which is the most important of multiple important TB-related findings on this image. There is no pleural effusion.

Partial correctness: Scarring in the right middle lobe, and right hilar and inferior mediastinal adenopathy may both allude to the large dense opacity in the medial lower right chest (probably consolidation in the right middle lobe related to TB).

Incorrectness: The small cavitary in the left upper lobe lesion related to TB is unaddressed.



(a) Best-ranked report in four-way comparison

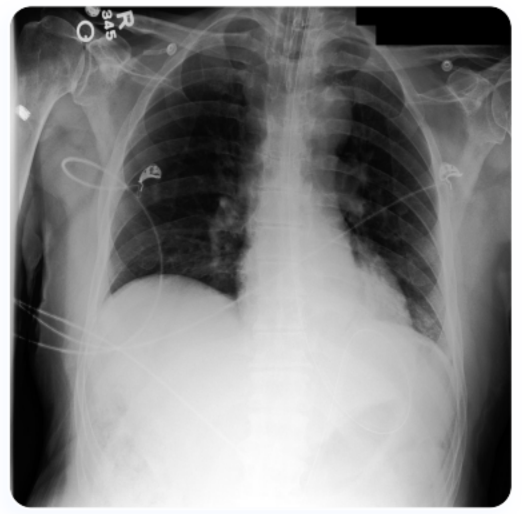
(b) Pairwise preference of each model scale compared to reference report

Indication

-year-old male with intubation. No prior examinations for comparison.

Reference Report

Endotracheal tube terminates 6 cm from the carina. Nasogastric tube loops within the stomach with side port within the gastric body. There are no significant pleural effusions or pneumothorax. Mild retrocardiac atelectasis is present. Heart size is normal. The aorta is tortuous.



Med-PaLM M 12B

An endotracheal tube terminates 5 cm above the carina. There is no pneumothorax or large pleural effusion. Opacities in the medial right lower lobe and retrocardiac region may represent atelectasis aspiration or pneumonia. A calcified lymph node is noted in the aortopulmonary window. Cardiac size is normal. *[No comment about NG/OG tube]*

Med-PaLM M 84B

Endotracheal tube terminates 5 cm above the carina. Nasogastric tube courses below the diaphragm and inferiorly beyond the film. There is retrocardiac opacity which may represent atelectasis or aspiration. The heart size is normal. The aorta is tortuous and unfolded. No pleural effusions or pneumothorax.

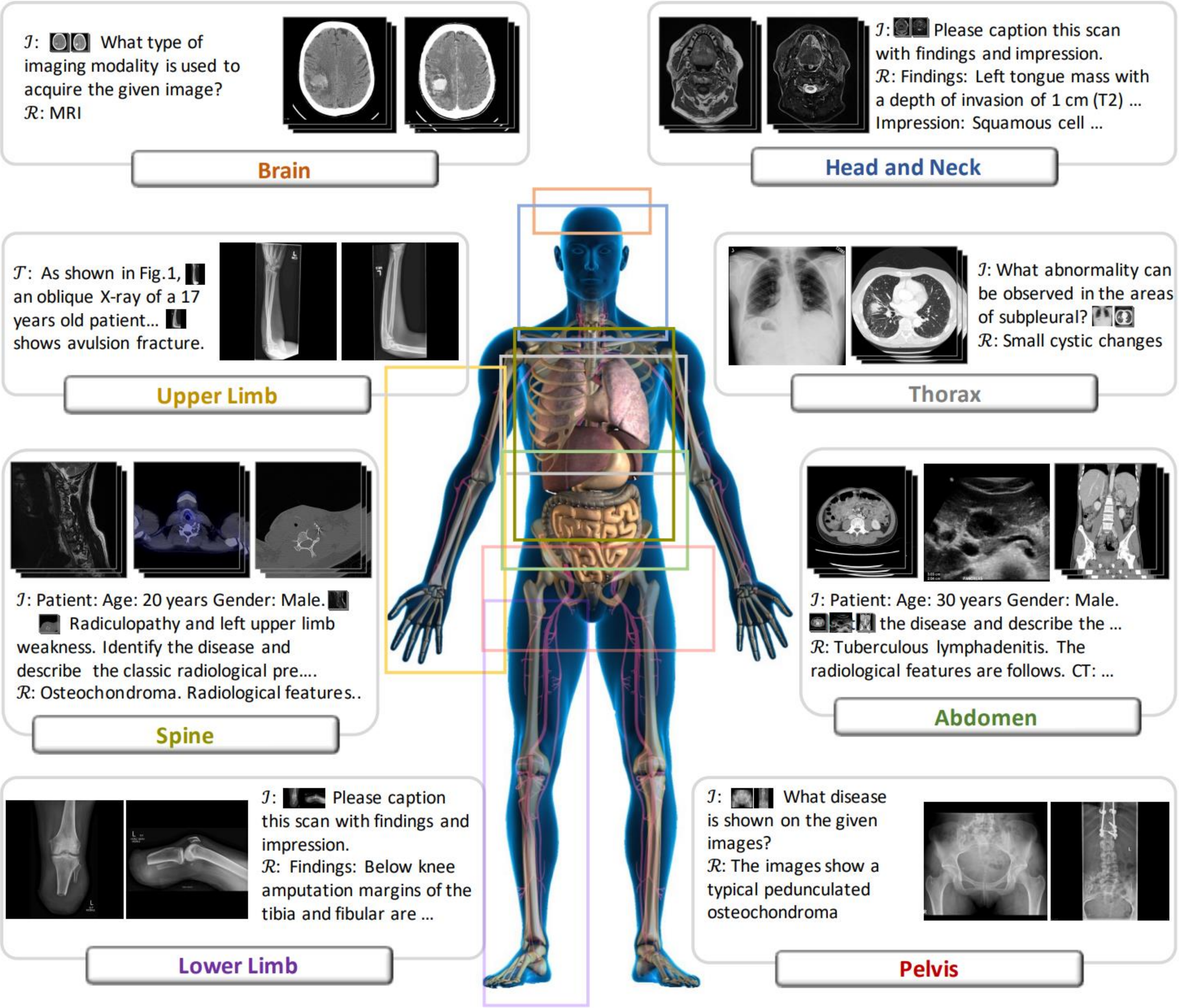
Med-PaLM M 562B

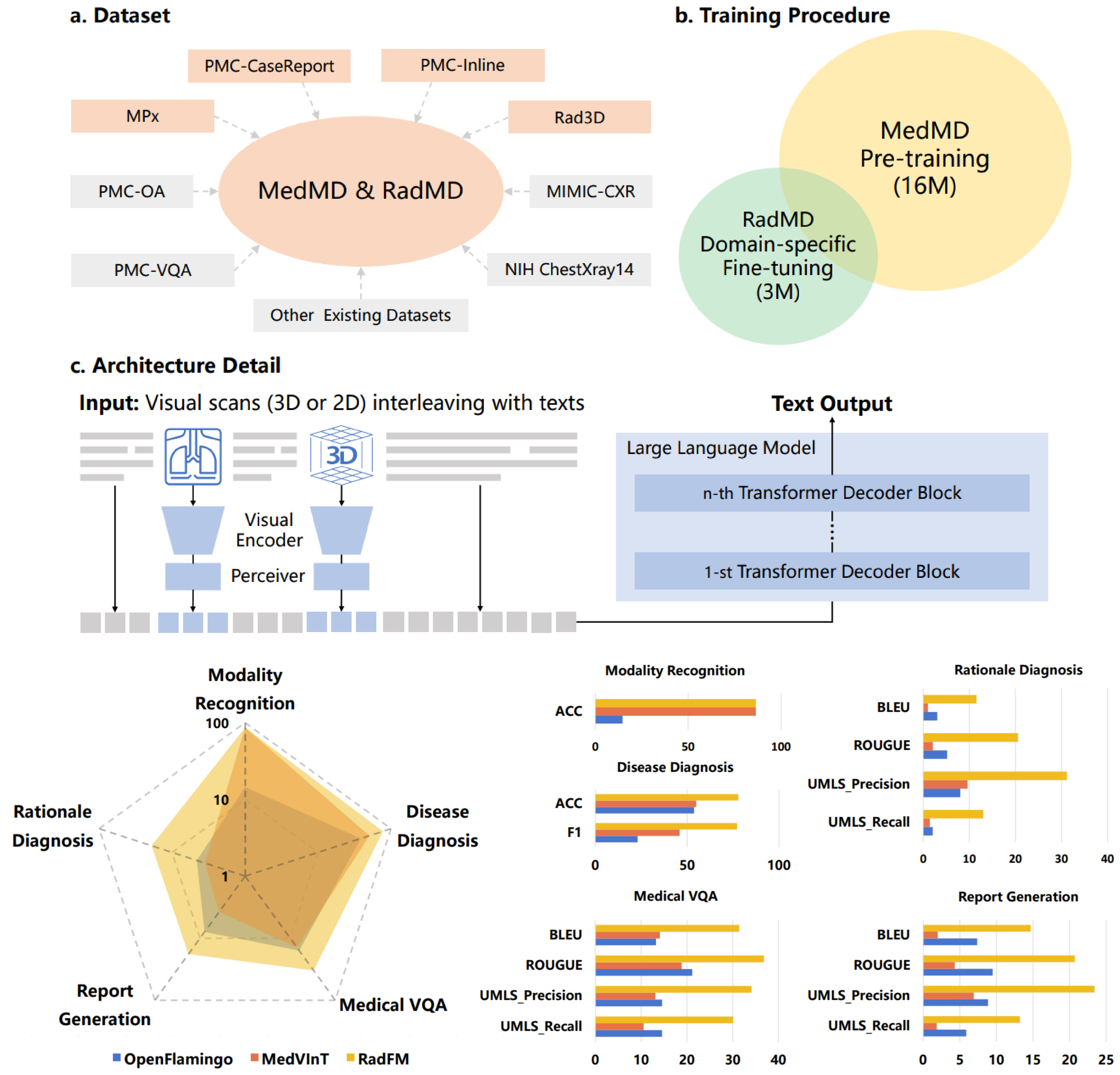
Endotracheal tube terminates 5 cm above the carina. Nasogastric tube courses below the diaphragm and inferiorly beyond the film. Lung volumes are low with crowding of bronchovascular markings. There is a left retrocardiac opacity. There is no pneumothorax or pleural effusions.

Significant error | Insignificant error | *[Omission]*

Source : Tao Tu et al., *arXiv2307.14334* (2023)

- **RadFM : Towards Generalist Foundation Model for Radiology**
  - Based on a large-scale Medical Multi-modal Dataset, MedMD, consisting of 16M 2D and 3D medical scans, Radiology Foundation Model, termed as RadFM is developed and evaluated.





Source : Chaoyi Wu et al., *arXiv2308.02463* (2023)

03

Regulatory and Ethical Considerations

▪ Challenges and Potential Risks of LLMs and LMMs as a Medical Device

- **Data Collection :** Clinical history, treatment outcome, genomic or other molecular data altogether?
- **Privacy :** Free from de-anonymization? Consent for multimodal data?

} **Development**

- **Regulatory Approval** : Medical Device or not? Approval of FM?
- **Clinical Trial** : How to evaluate FM-based medical devices or GMAI?

} **Validation**

- **Usability :** What would be the user interface of GMAI? Input and output? Jailbreak-free?
- **Deployment** : Cloud-based or on-premise. Multimodal Integration? Continuously updated or frozen models.

} **Deployment**

- **Safety** : Hallucination-free? Trustworthiness of the output? Over or mis-use?
- **Liability** : Who's in charge of adverse event? FM developer? Service Provider? Clinician?

} **Monitoring**

Table 1 | Challenges in the regulatory approval of large language models

Challenge	Details
Verification	Near-infinite range of inputs and outputs, including hallucinated outputs, make these models untestable
Provenance	No control over provenance when used as an underlying model on which a medical device is built
Changes	Not a fixed model, as the generative approaches and the manual and automated constraining of outputs (for example, to limit harmful advice) can be adapted on market
Usability	Near-infinite range of user experiences, depending on the input
Risks	No proven method to prevent harmful outputs
Surveillance	A near-infinite number of outputs make surveillance impossible

Table 3. A list of regulatory challenges related to the rise of LLMs.

Regulatory challenge	Short description
Patient Data Privacy	Ensuring that patient data used for training large language models are fully anonymized and protected from potential breaches. This poses a significant regulatory challenge, as any violation could lead to serious consequences under privacy laws like HIPAA in the US.
Intellectual Property	If an LLM generates content similar to proprietary medical research or literature, it could lead to issues regarding intellectual property rights.
Medical Malpractice Liability	Determining who is responsible when an AI's recommendations lead to patient harm. Is it the AI developers, the healthcare professionals who used it, or the institutions that adopted it?
Quality Control & Standardization	Regulation is required to ensure the reliability and consistency of AI-generated medical advice, which can vary based on the data used to train the AI.
Informed Consent	Patients need to be informed and give consent when AI tools are used in their healthcare management. This is challenging because it can be difficult for patients to fully understand the implications of AI use.
Interpretability & Transparency	Regulations need to ensure transparency about how decisions are made by the AI. This is particularly challenging with AI models that are often termed as "black boxes" due to their complex algorithms.
Fairness and Bias	Regulation is needed to prevent biases in AI models, which could be introduced during the training process using patient data. This can lead to disparities in healthcare outcomes.
Data Ownership	It can be challenging to define and regulate who owns the data that large language models learn from, especially when it comes to patient data.
Over-reliance on AI Models	Over-reliance on AI could lead to decreased human expertise and potential errors if the AI malfunctions or provides incorrect information. Regulations are needed to balance the use of AI and human expertise.
Continuous Monitoring & Validation	Ensuring the continuous performance, accuracy, and validity of AI tools over time and across different populations is a critical regulatory challenge.

S. Gilbert et al., *Nature Medicine*(2023), B. Mesko et al., *npj Digital Medicine*(2023)

- **Can LLMs be regulated as a Software as a Medical Device?**

Comment

https://doi.org/10.1038/s41591-023-02412-6

Large language model AI chatbots require approval as medical devices

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt & Paul Wicks

Check for updates

Chatbots powered by artificial intelligence used in patient care are regulated as medical devices, but their unreliability precludes approval as such.

Every new technology must satisfy concerns of safety, performance and risk/benefit to flourish. Large language models (LLMs) are neural network language models that include OpenAI's Generative pre-trained transformer (GPT) and Google's Pathways Language Model (PaLM)[1,2]. ChatGPT is an LLM chatbot launched in November 2022 that has remarkable conversational capability and the capacity to near-instantly and creatively mimic different human conversational styles on request[2,3]. Its underlying LLM was updated from GPT-3.5 to GPT-4 in March 2023[3]. It has been proposed that LLM chatbots can be applied in medicine, in which information exchange, advice and the linking of information flows are crucial parts of service delivery[3,4]. Today, however, developers of LLM chatbots acknowledge that they often generate highly convincing statements that are verifiably wrong, as well as sometimes 'hallucinating' information or providing inappropriate responses to questions[2,3] (Table 1).

LLM chatbots produce a 'reasonable continuation' of text, starting from a prompt, using their tokenized encoding of language extracted from billions of unidentified general web pages and books[5]. Their development includes both non-supervised and supervised learning, with numerous trial-and-error human decisions to optimize their plausibility and reasonableness. Today there is no way to be certain about the quality, evidence level or consistency of clinical information or supporting evidence for any LLM response. LLMs simply reassemble what was most commonly written by humans[5], and when asked to produce a source, they will frequently invent a plausible, but nonexistent, citation[6].

Regulation as medical devices

Some claim that these tools could transform medicine through applications including streamlining communication tasks and providing clinical decision support (CDS)[3,4]. However, under EU and US law, tools intended for most of these use cases must be approved as medical devices (Table 2). Software performing anything more than simple database functions to assist in diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease is categorized as a medical device, and regulatory controls apply, including the requirement that the tools be developed in a quality-management system. The EU also requires post-market surveillance and clinical follow-up, which are particularly challenging for LLMs. As they have no inheritable quality assurance from their developers, they are excluded from use as external 'plug-in' components of medical devices, such as through an application programming interface (API).

Table 1 | Challenges in the regulatory approval of large language models

Challenge	Details
Verification	Near-infinite range of inputs and outputs, including hallucinated outputs, make these models untestable
Provenance	No control over provenance when used as an underlying model on which a medical device is built
Changes	Not a fixed model, as the generative approaches and the manual and automated constraining of outputs (for example, to limit harmful advice) can be adapted on market
Usability	Near-infinite range of user experiences, depending on the input
Risks	No proven method to prevent harmful outputs
Surveillance	A near-infinite number of outputs make surveillance impossible

Because they have a near-infinite range of inputs and outputs, it is challenging to test the usability and on-market performance of LLMs, and so it is questionable whether their tendency to suggest harmful or false – yet highly plausible – information can ever be controlled (Table 1). In their current state, LLMs do not ask for missing information needed to provide an accurate answer, provide no accompanying indication of relative certainty or confidence, and generally provide no genuine sources. This rules out their use in the USA for non-device clinical decision support. It also makes it extremely challenging to verify the outputs of the design process, mitigate all identified risks and demonstrate valid clinical association between inputs and outputs, both prior to approval and in ongoing monitoring after their market release; these issues effectively rule out their valid marketing as medical devices under current EU law.

Arguably, search engines have transformed medicine, and despite protests from medical providers, some two-thirds of patients carry out an online search on their symptoms before a medical consultation[7]; most doctors do so too, checking search engines between one and three times a day[8]. Search engines have roles in decision-making, but this does not make them regulated medical devices, as their developers did not have the intended purpose of providing a tool for medical diagnosis, decision support or planning therapy when creating them[9,10].

The forthcoming integration of LLM chatbots into search engines may increase users' confidence in the search results by mimicking conversation[11]. However, it has been demonstrated that LLMs can provide profoundly dangerous information when prompted with medical questions[2]. Alarming examples of this include an LLM advising a simulated mental health patient to kill themselves[12] and a version of Bing occasionally threatening users. Nonetheless, at least one mental health counselling app has used GPT-3 to respond to users, without

Table 3 | Steps for the approval of LLM-enabled medical devices

Regulatory requirement and challenge	Possible approaches for developers
A clearly defined intended purpose	Limit responses to user prompts to the stated scope of the device. Limit scope and severity to exclude emergency or critical use.
Determine the risk class according to national frameworks	Design to inform, not drive, medical decisions.
Define requirements	• Implement performance benchmarks for narrow use cases. Stop or highly constrain on-market adaptivity, and document this. • Constrain the LLM to stop harmful advice and improve safety. • Control risks of data leakage and unconsented use of data to comply with data protection. • Provide accessible information to users on safe use.
Demonstrated analytical and clinical validity of output to inputs and use validatable knowledge base	• Use only an LLM self-developed by the medical device manufacturer or an LLM developed by an external provider that is documented for use in a medical device. • Develop from authoritative medical sources. • Rigorously test, constrain, retest and document. • Link automated real-time fact checking approaches in feedback to the LLM. • Carry out comprehensive clinical trials to demonstrate the safety and efficacy of the LLM medical device, following regulations and reporting frameworks.

S. Gilbert et al., *Nature Medicine*(2023)

▪ International Efforts for Guiding Proper Use of LLM and LMM in Healthcare

Correspondence

https://doi.org/10.1038/s41591-023-02656-2

Reporting standards for the use of large language model-linked chatbots for health advice

Check for updates

Modern artificial intelligence (AI)-based chatbots such as ChatGPT leverage the technology of large language models (LLMs) and natural language processing[1]. These chatbots use natural language processing to enable user interactions via a chat-based interface that produces LLM-generated text. Their performance has improved as LLM-linked chatbots are increasingly being trained with online sources, and their impressive qualities have garnered interest in their application for healthcare purposes[1].

Physicians and patients regularly use the internet for health advice[2,3], and the wide availability of LLM-linked chatbots has raised concerns for patient safety due to the accuracy of their output[4]. Healthcare professionals have a fiduciary duty in supporting patients to access accurate information, while avoiding false or misleading sources. For this reason, there has been substantial research interest in investigating whether LLM-linked chatbots can provide accurate clinical advice. However, the reporting of the design, conduct and analysis of this growing body of research is inconsistent.

Studies have assessed the performance of LLM-linked chatbots in providing a range of advice for both clinicians and patients. These have included topics such as health prevention[5], screening[6], diagnosis[7], differential diagnosis[8] and treatment[9]. Unlike other types of biomedical research, no reporting standards exist for studies that evaluate the performance of LLM-linked chatbots when providing clinical advice. It is well documented that the lack of consistent and transparent reporting of research prevents readers from appraising the study methods and obfuscates the interpretation of their findings[10]. Owing to the rapidly expanding literature in this space, there is a need for structured guidance and minimum reporting recommendations for researchers interested in evaluating the performance of LLM-linked chatbots in summarizing health evidence and providing clinical advice.

To develop a reporting guideline for chatbot assessment studies, we have gathered a diverse group of international stakeholders including statisticians, research methodologists, reporting guideline developers, natural language processing researchers, journal editors, chatbot researchers and patient partners. The development of the chatbot assessment reporting tool (CHART) is registered with the EQUATOR (enhancing the quality and transparency of health research) international network. This guideline will generate a reporting checklist and flow diagram by adhering to robust methodology[10], as well as the evidence-based EQUATOR toolkit on developing reporting guidelines. This will involve a review of the literature to identify relevant reporting guidelines and chatbot assessment studies. Next, a Delphi consensus will be conducted among our stakeholders with public engagement prior to focused, synchronous consensus meetings. Researchers, clinicians and patients will benefit from the establishment of structured reporting standards to encourage transparent reporting among chatbot assessment studies. We invite motivated stakeholders to participate in this collaborative, multidisciplinary initiative.

Bright Huo[1], **Giovanni E. Cacciamani**[2,3], **Gary S. Collins**[4,5], **Tyler McKechnie**[1,6], **Yung Lee**[1,7] **& Gordon Guyatt**[6,8]
[1]Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Ontario, Canada. [2]USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [3]AI Center at USC Urology, University of Southern California, Los Angeles, CA, USA. [4]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK. [5]UK EQUATOR Centre, University of Oxford, Oxford, UK. [6]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada. [7]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [8]Department of Medicine, McMaster University, Hamilton, Ontario, Canada.
e-mail: brighthuo@dal.ca

Published online: 13 November 2023

References

1. Thirunavukarasu, A. J. et al. *Nat. Med.* **29**, 1930–1940 (2023).
2. Amante, D. J. et al. *J. Med. Internet Res.* **7**, e106 (2015).
3. Wang, C. et al. *Med. Internet Res.* **25**, e48009 (2023).
4. Mikalef, P. et al. *Health Info. Libr. J.* **34**, 58–73 (2017).
5. Lee, T. C. et al. *Gastroenterology.* **165**, 509–511 (2023).
6. Rahsepar, A. A. et al. *Radiology* **307**, e230922 (2023).
7. Emilie, S. H. et al. *Surgery* **174**, 1273–1275 (2023).
8. Qu, R. W. et al. *OTO Open.* **7**, e67 (2023).
9. Zhou, Z. et al. *Eur Urol.* **84**, 355–356 (2023).
10. Moher, D. et al. *PLoS Med.* **7**, e1000217 (2010).

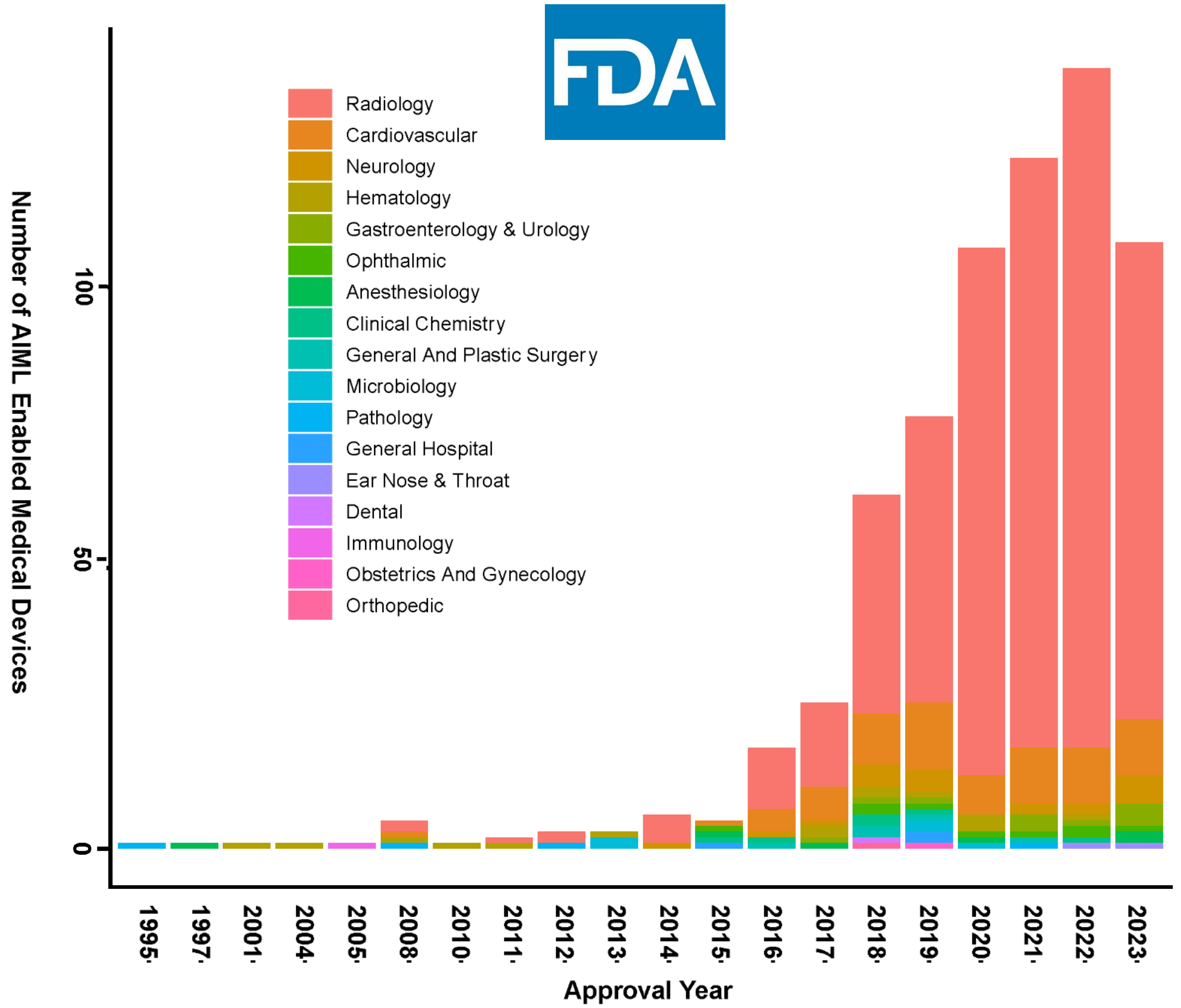
Competing interests

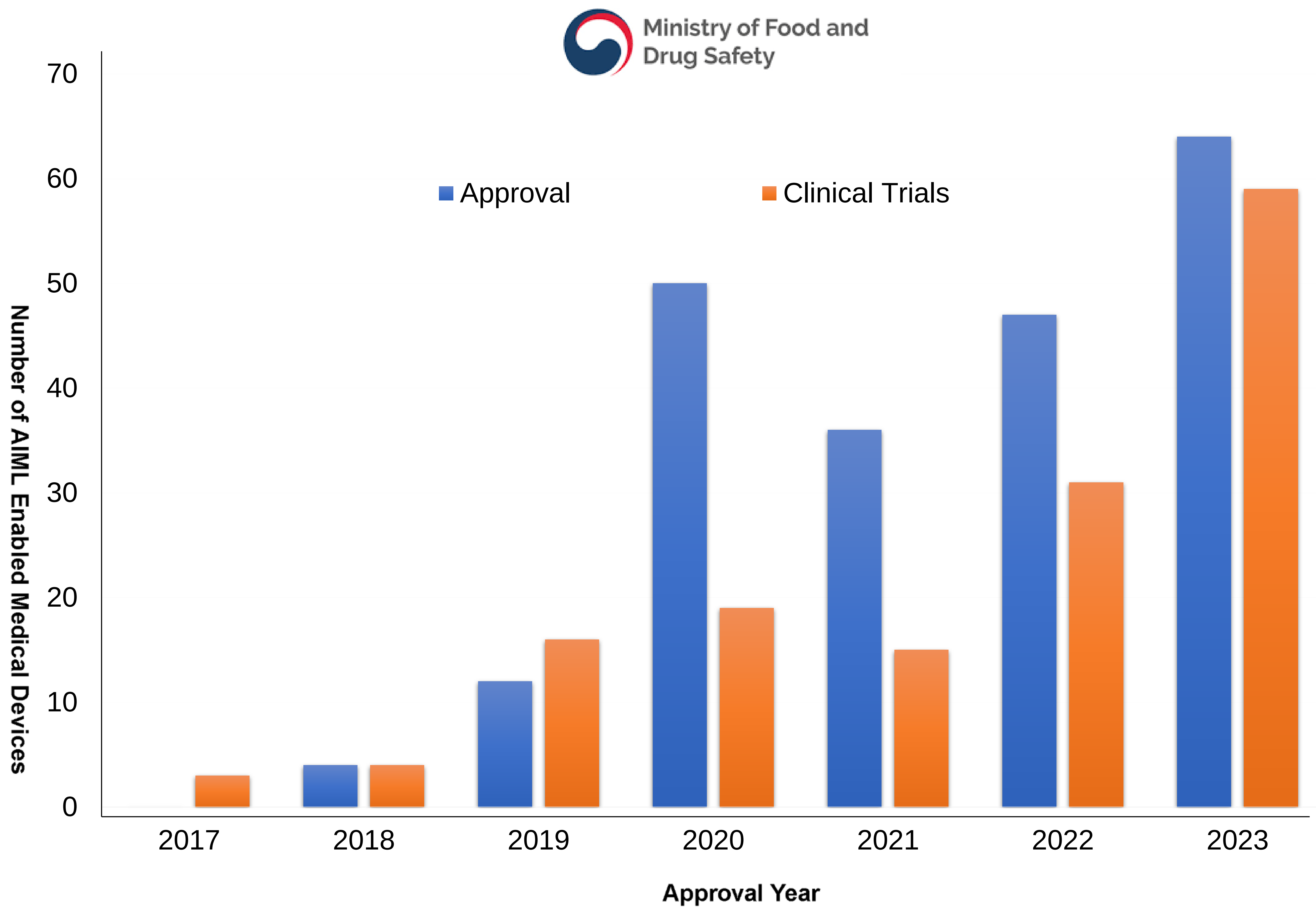
The authors declare no competing interests.

WHO(2024), B. Huo et al., *Nature Medicine*(2023)

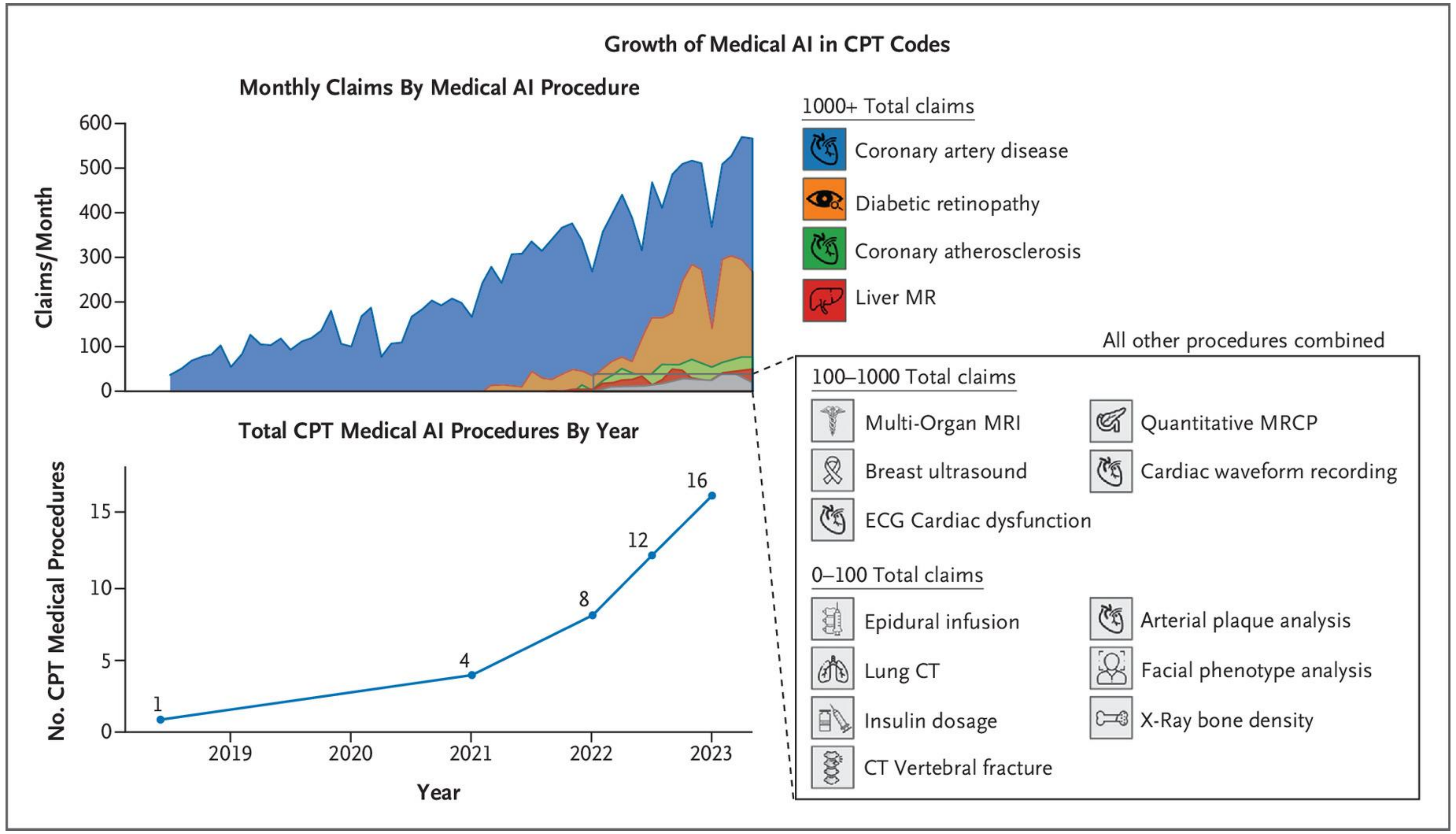
Current Status of AI-based SaMD

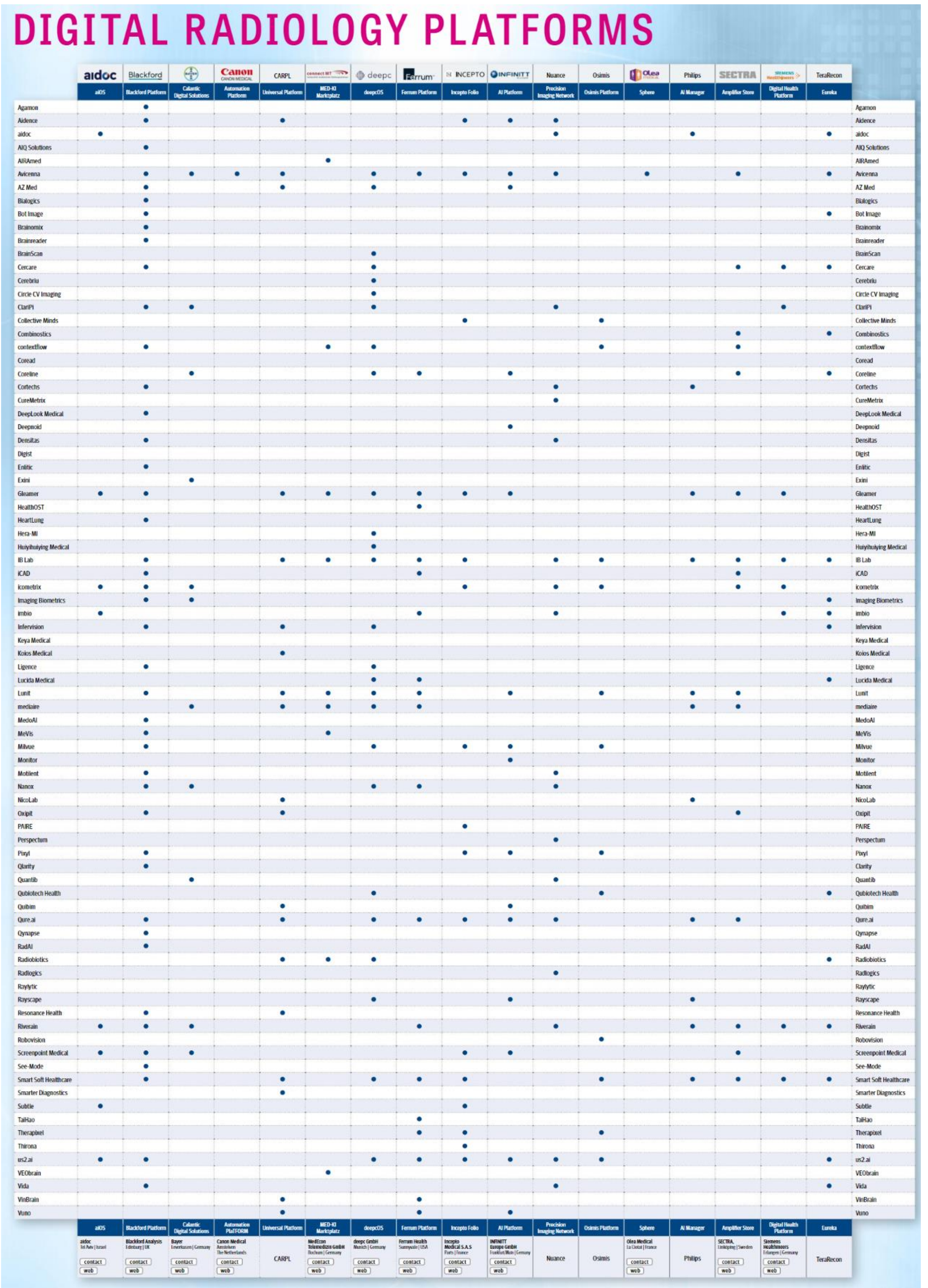
- The number of FDA/MFDS approved/cleared AI/ML-based SaMD is increasing fast.
- However, most of the approved medical devices are using unimodal input, intended for specific use.





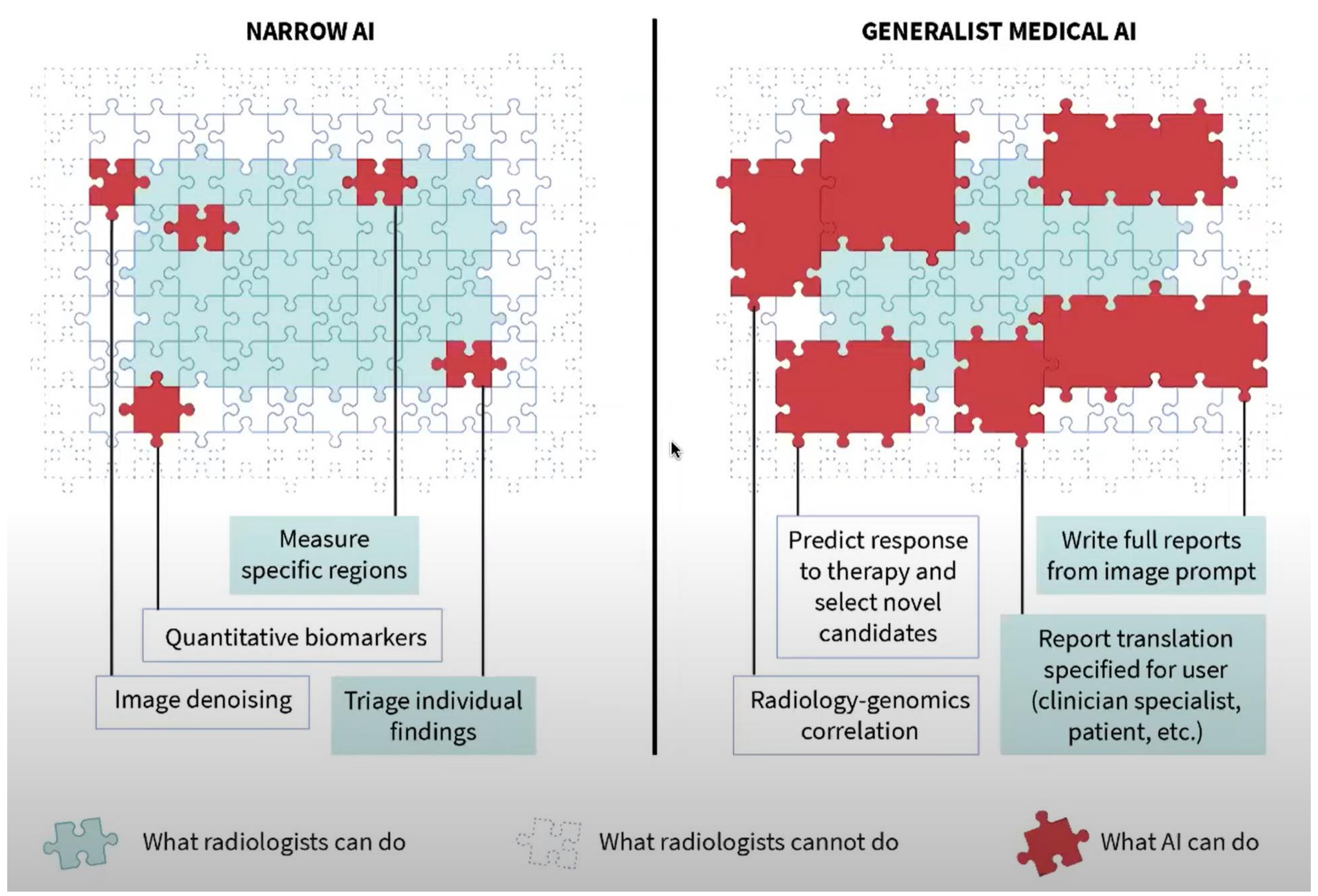
- **Current Status of AI-based SaMD**
  - Even though the number of claims are increasing, AI/ML adoption is slower than expected
  - This is partly due to the fragmented product/market structure of AI/ML SaMD

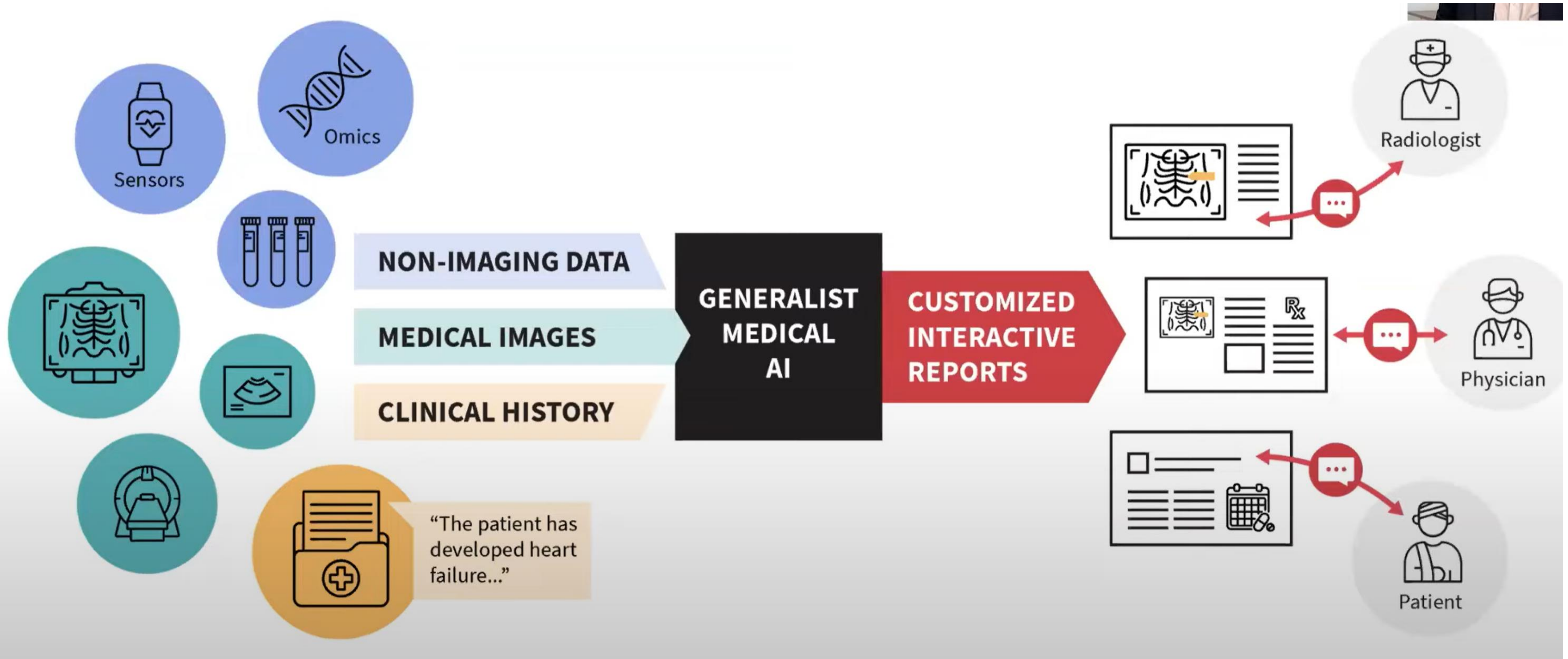




Kevin Wu et al., *NEJM AI*(2024),https://dieurope.com/radiology-ai-at-a-glance/

- **Toward Generalist Medical AI**
  - A generalist medical AI(GMAI) models will be capable of carrying out a diverse set of tasks using very little or no task-specific labelled data.
  - Built through self-supervision on large, diverse datasets, GMAI will flexibly interpret different combinations of medical modalities, including data from imaging, electronic health records, laboratory results, genomics, graphs or medical text.
  - Models will in turn produce expressive outputs such as free-text explanations, spoken recommendations or image annotations that demonstrate advanced medical reasoning abilities.





Pranav Rajpurkar (2023)

Thank you!

Kyu-Hwan Jung, Ph.D.
(06351) Rm5, B1, Research Institute for Future Medicine,
Irwon-ro 81, Gangnam-gu, Seoul, South Korea

T +82.10.8562.3820
E khwanjung@skku.edu