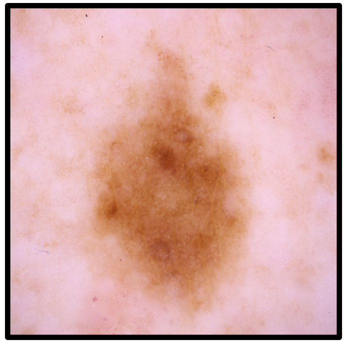




Language Guided Concept Bottlenecks for Interpretable and Robust Image Classification

Yue Yang

WPE-II Presentation

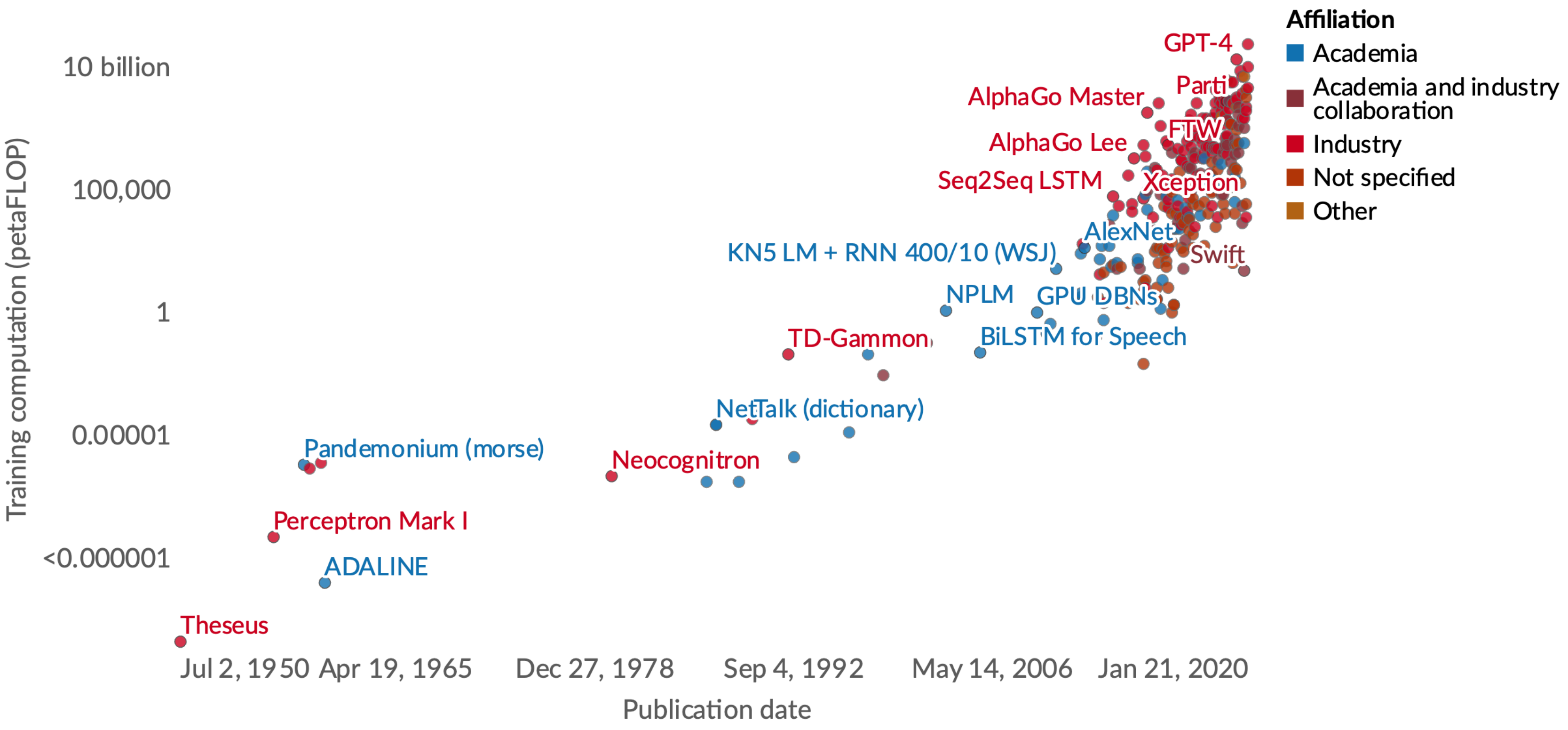
Committee: Dan Roth (Chair), Chris Callison-Burch, Mark Yatskar

Models are getting performant but less interpretable.

Computation used to train notable AI systems, by affiliation of researchers

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



Data source: Epoch (2023)

OurWorldInData.org/artificial-intelligence |

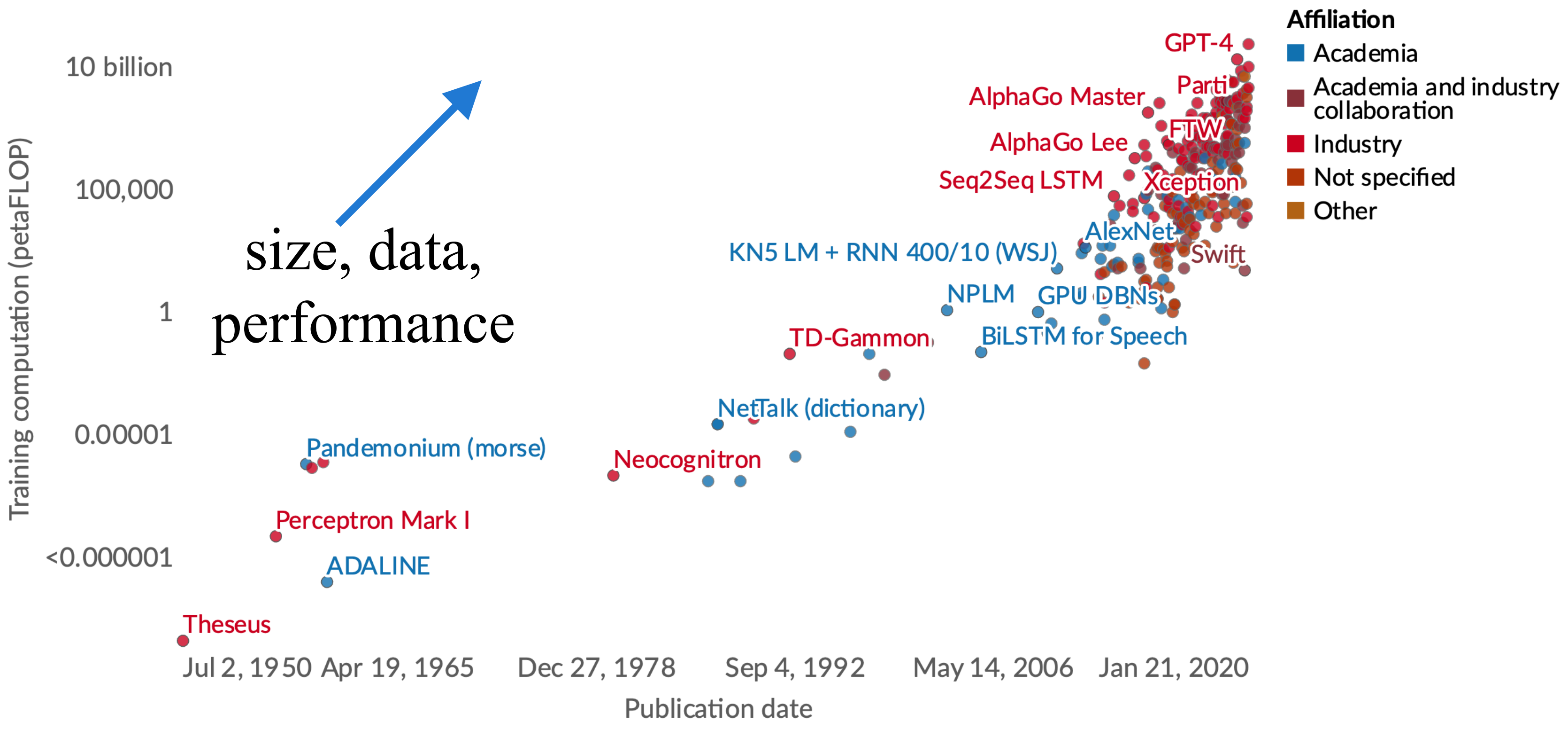
Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

Models are getting performant but less interpretable.

Computation used to train notable AI systems, by affiliation of researchers

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.

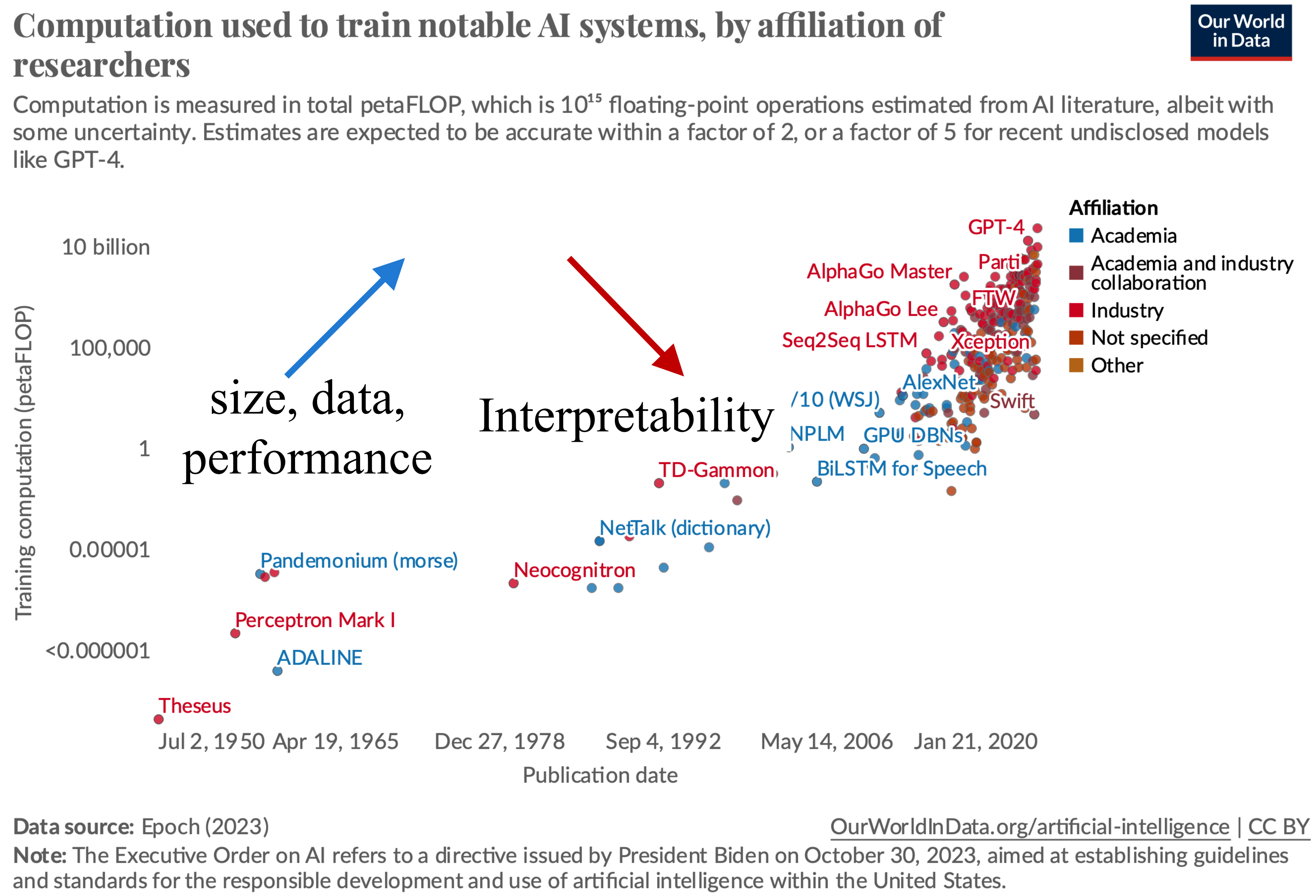


Data source: Epoch (2023)

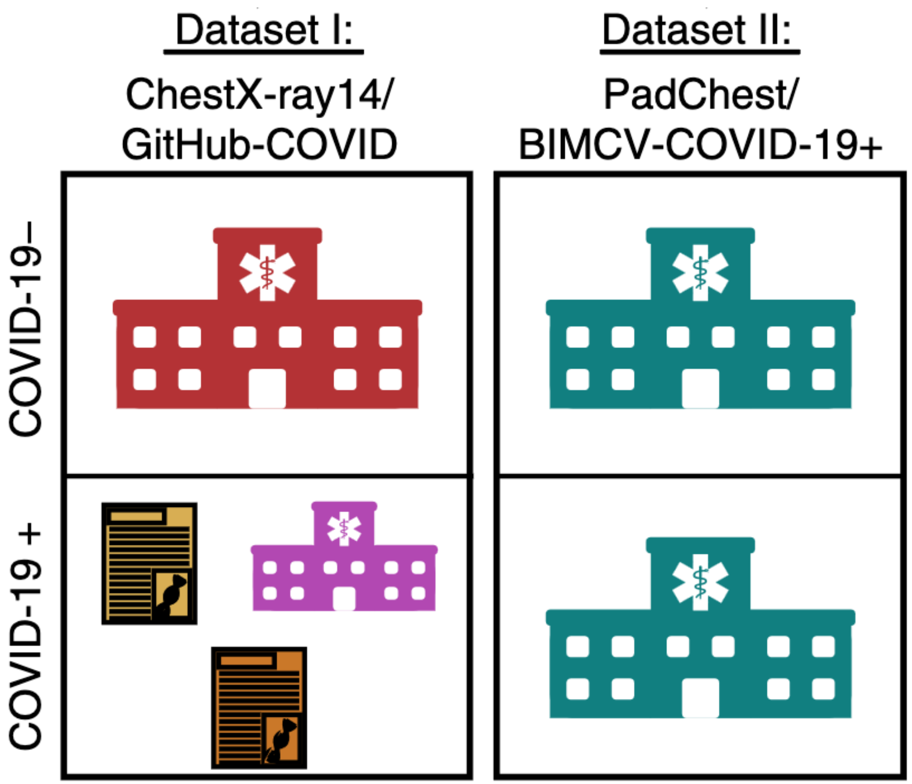
OurWorldInData.org/artificial-intelligence |

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

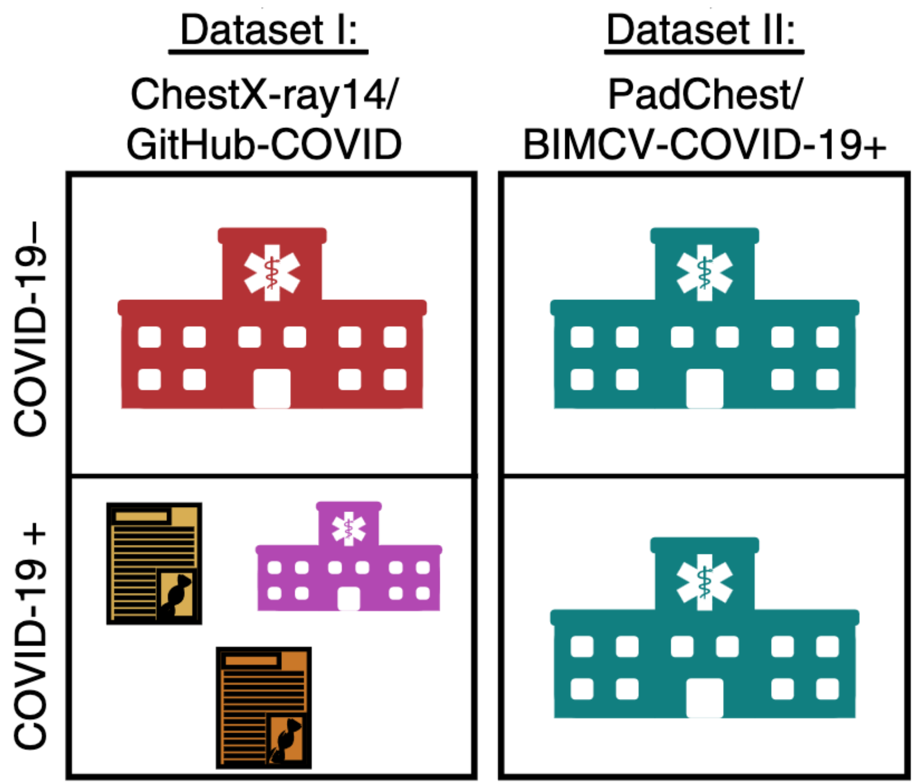
Models are getting performant but less interpretable.



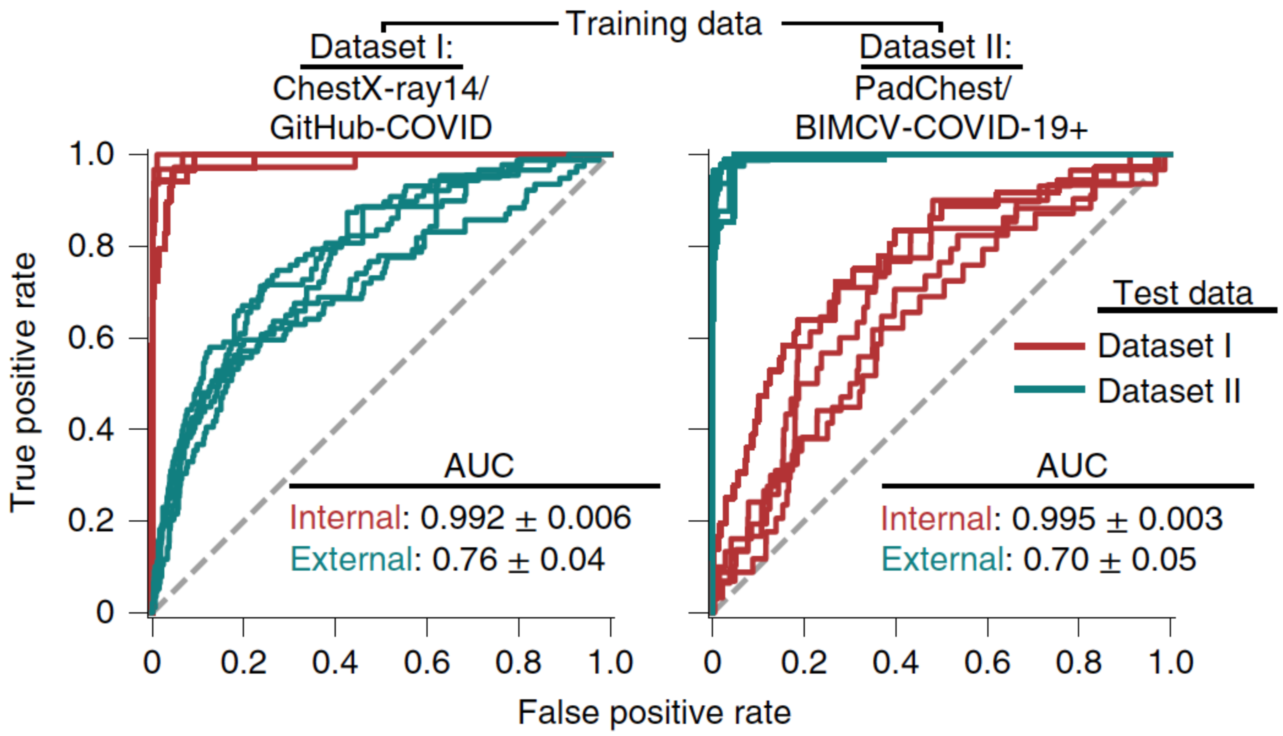
Catastrophic failures in critical domains.



Catastrophic failures in critical domains.



Poor generalization.

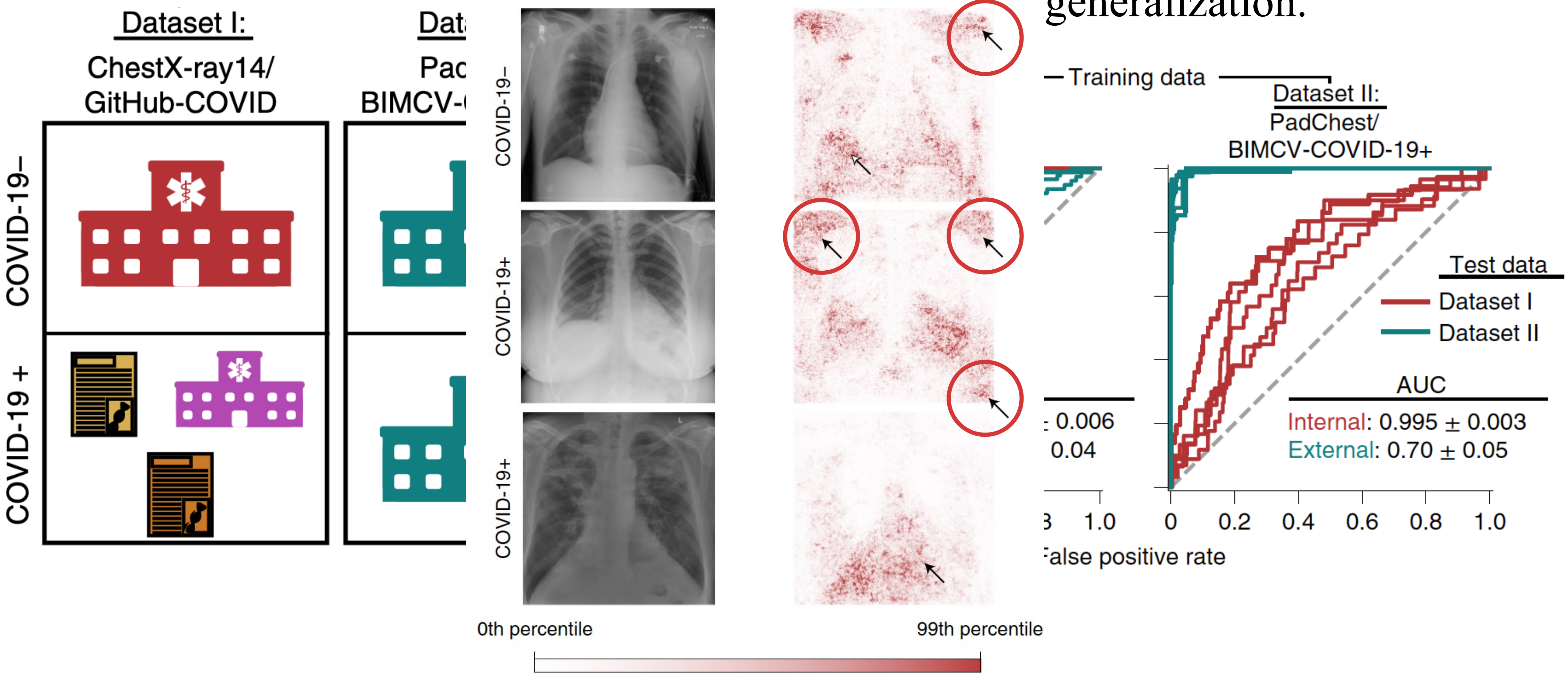


DeGrave et al. Nature Machine Intelligence. 2023

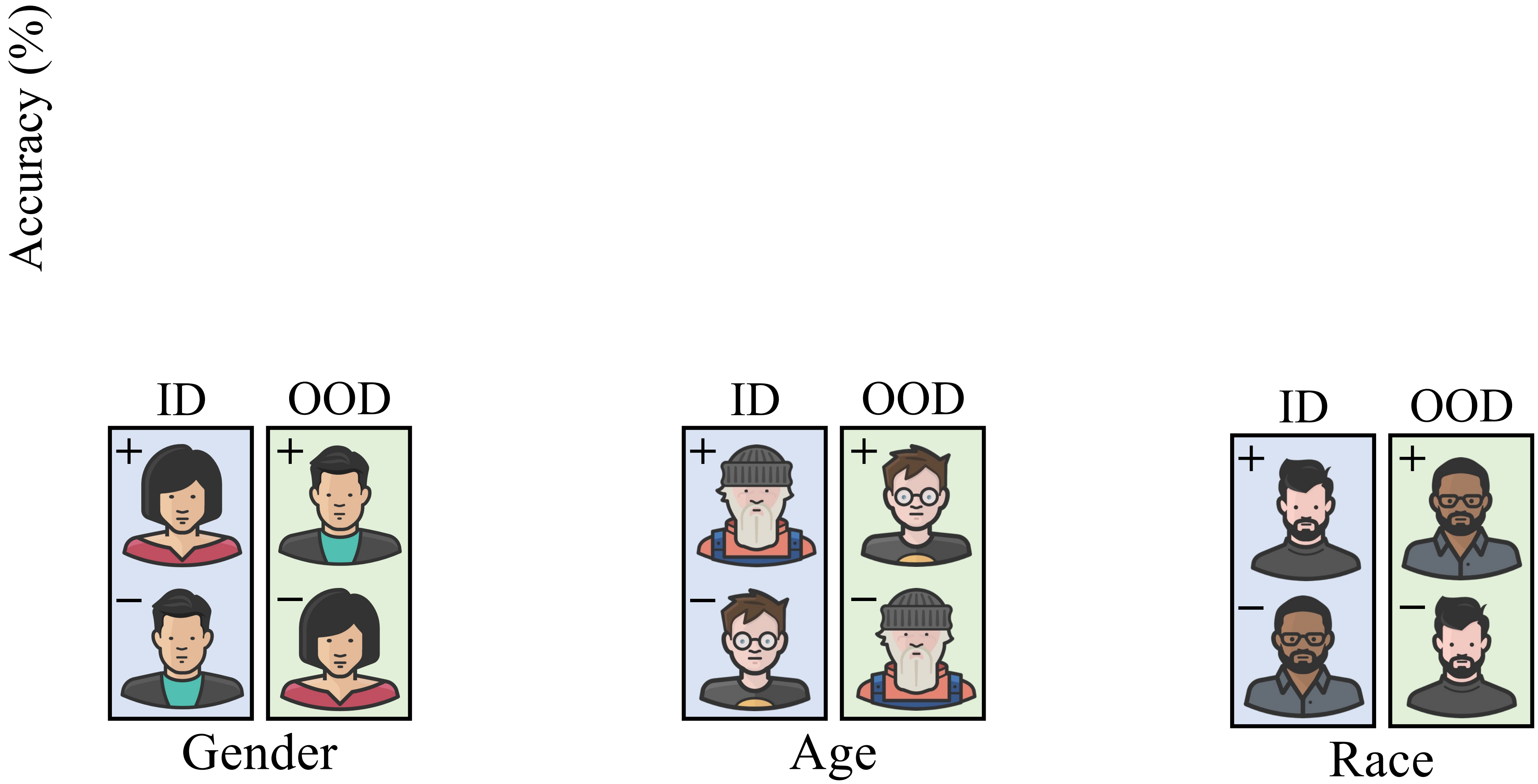
Catastrophic failures in critical domains.

Select shortcuts.

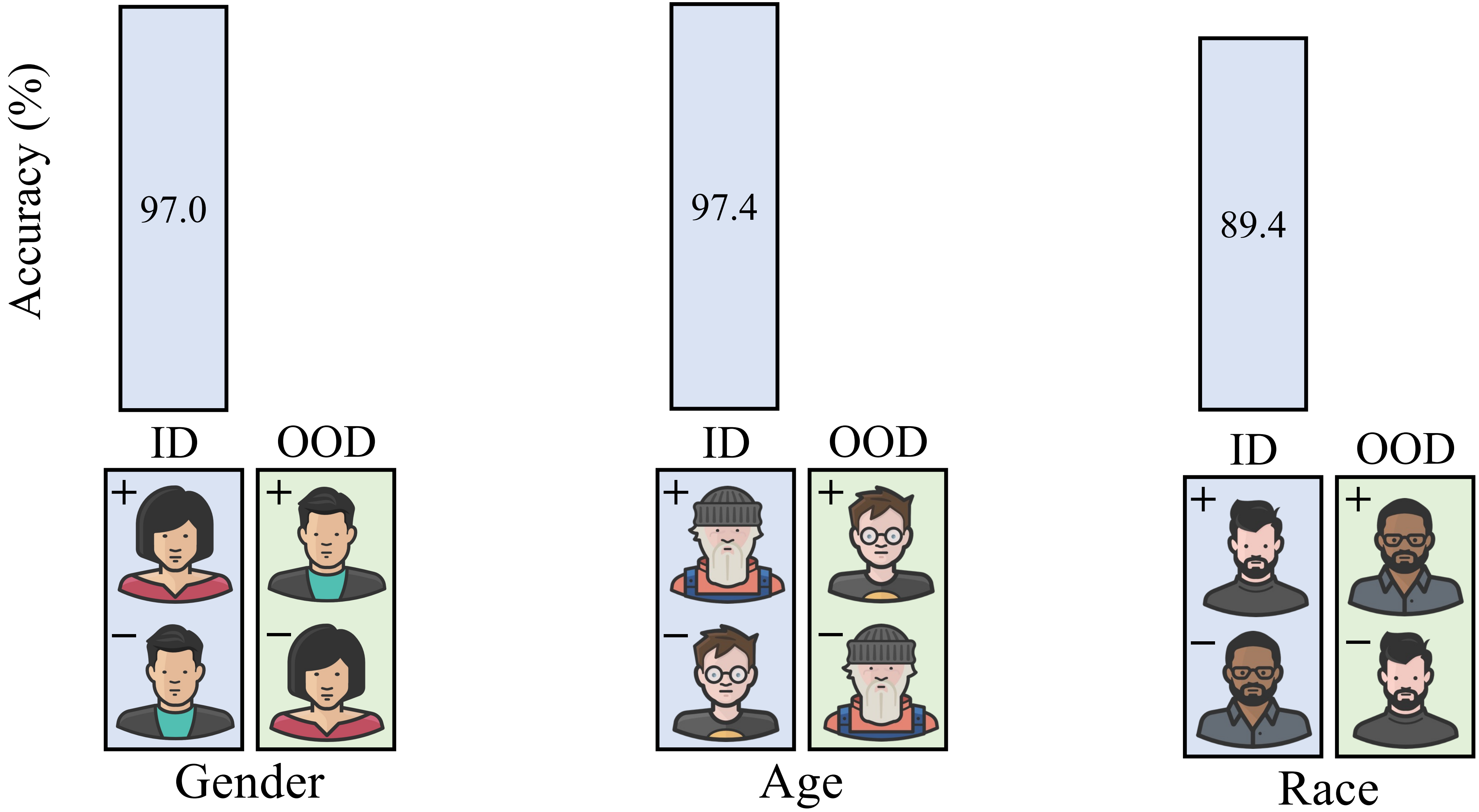
Poor generalization.



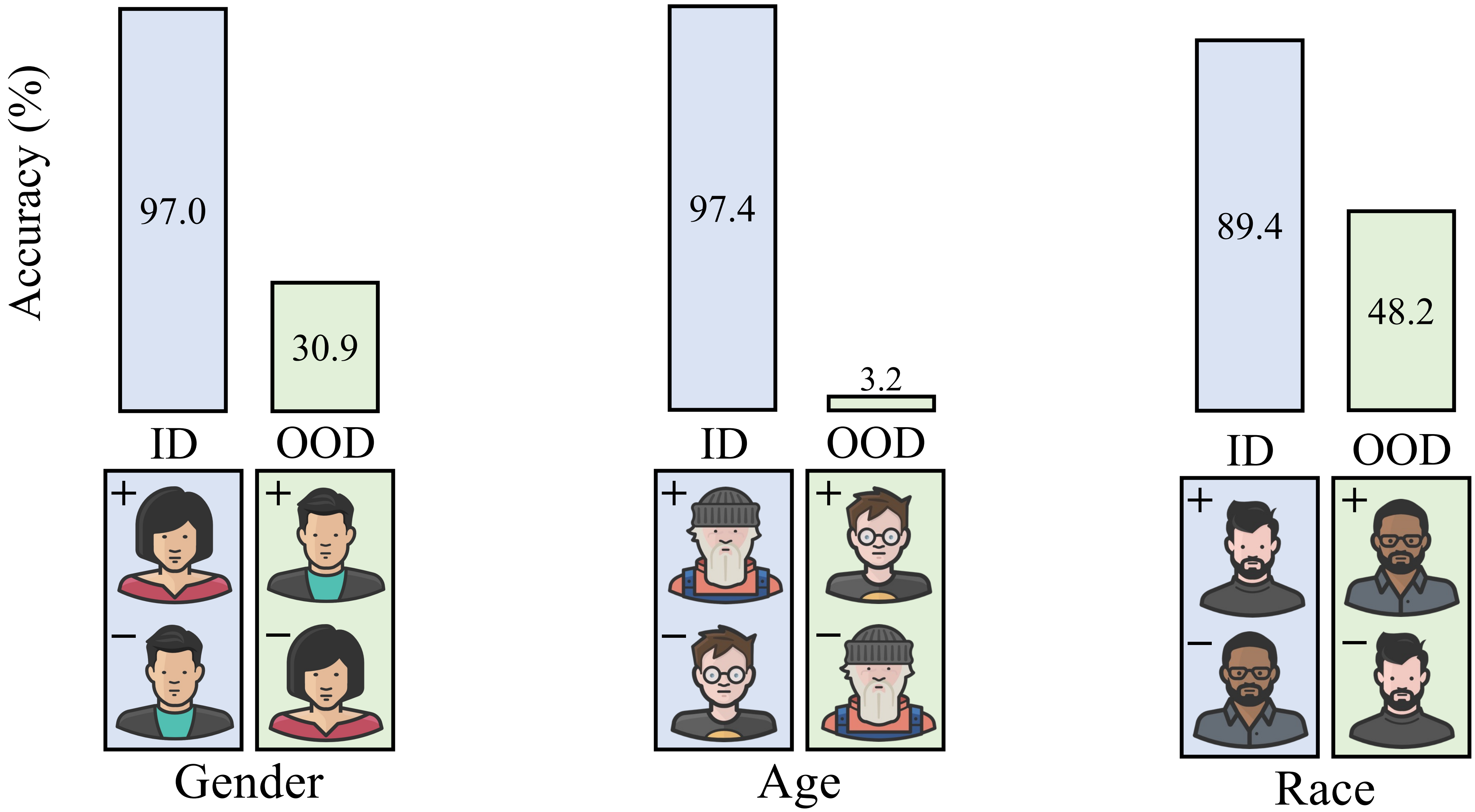
Black-box models generalize poorly on domain shifts.



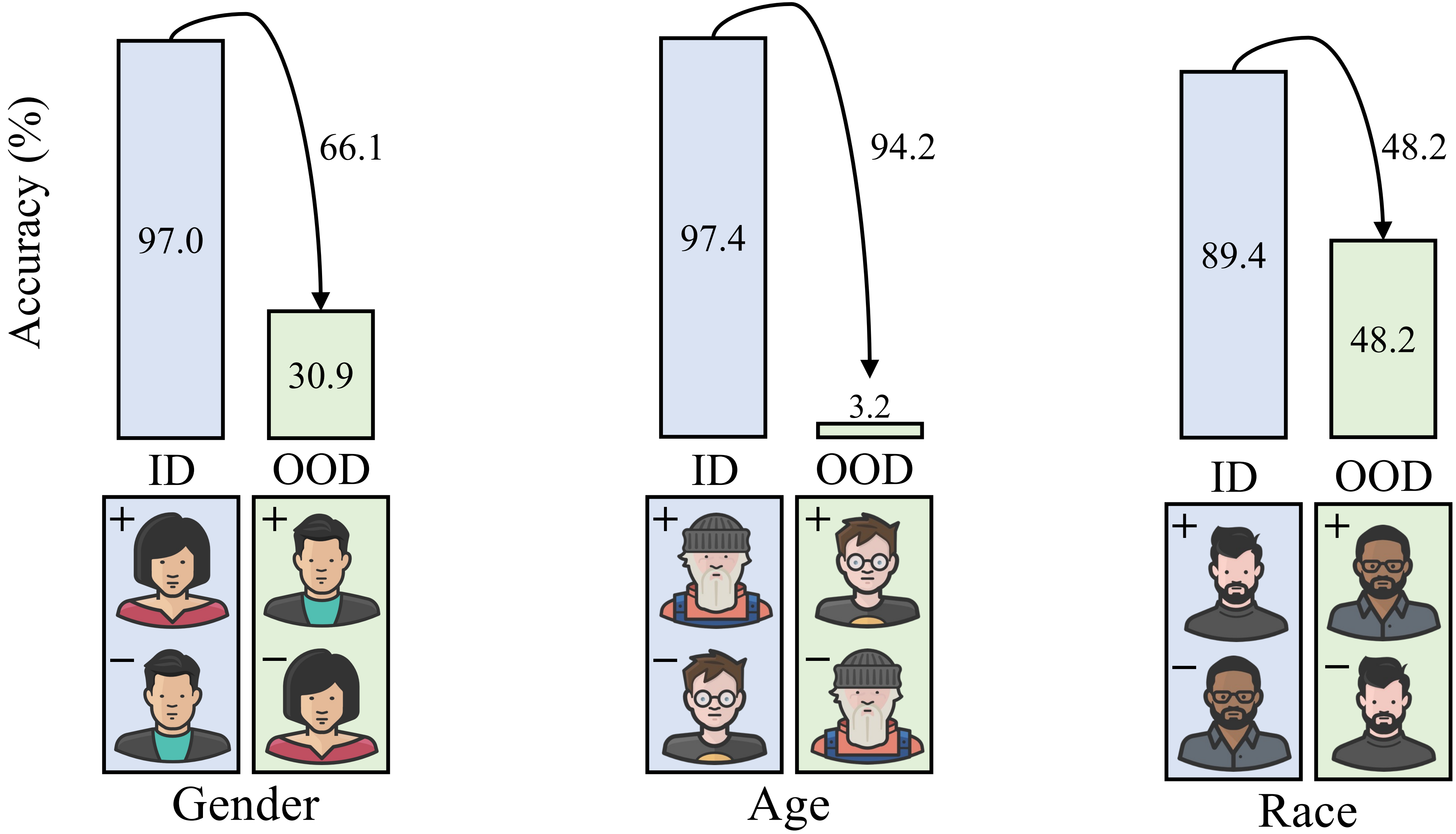
Black-box models generalize poorly on domain shifts.



Black-box models generalize poorly on domain shifts.



Black-box models generalize poorly on domain shifts.



Solution-1: Post-hoc Explanation

- Explain the black box model with another black box model.
- Explanations are often **not faithful** and can be misleading.

	Test Image
Explanations Using Attention Maps	

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019.

Solution-1: Post-hoc Explanation

- Explain the black box model with another black box model.
- Explanations are often **not faithful** and can be misleading.

	Test Image	Evidence for Animal Being a Siberian Husky
Explanations Using Attention Maps		

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019.

Solution-1: Post-hoc Explanation

- Explain the black box model with another black box model.
- Explanations are often **not faithful** and can be misleading.

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019.

Solution-2: Inherently Interpretable Methods

Provide their **own explanations** that are **faithful** to the predictions.

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	⋯
2.	Prior arrests ≥ 5	1 point	+⋯
3.	Prior arrests for local ordinance	1 point	+⋯
4.	Age at release between 18 to 24	1 point	+⋯
5.	Age at release ≥ 40	−1 point	+⋯
		Score	= ⋯

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

Rudin and Ustun. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. INFORMS. 2018

Concept Bottleneck Models (CBMs)

Input Image *x*

End-to-end Model

label *y* (black-throated sparrow)

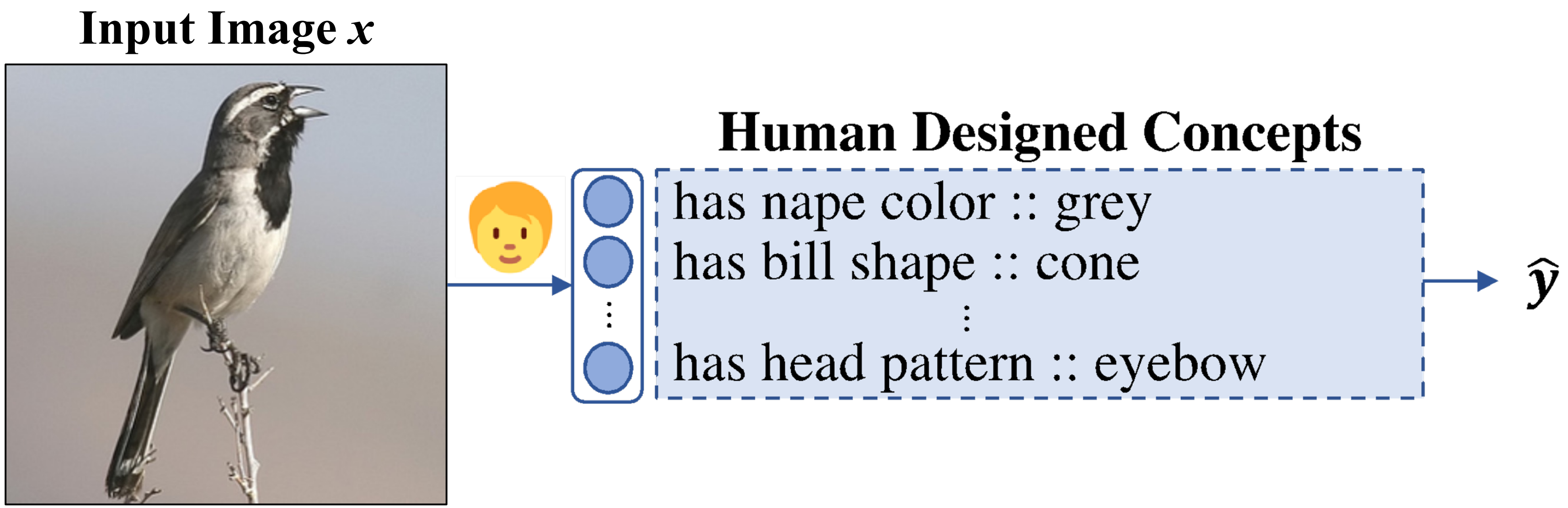
Concept Bottleneck Models (CBMs)

Input Image *x*

Black Box

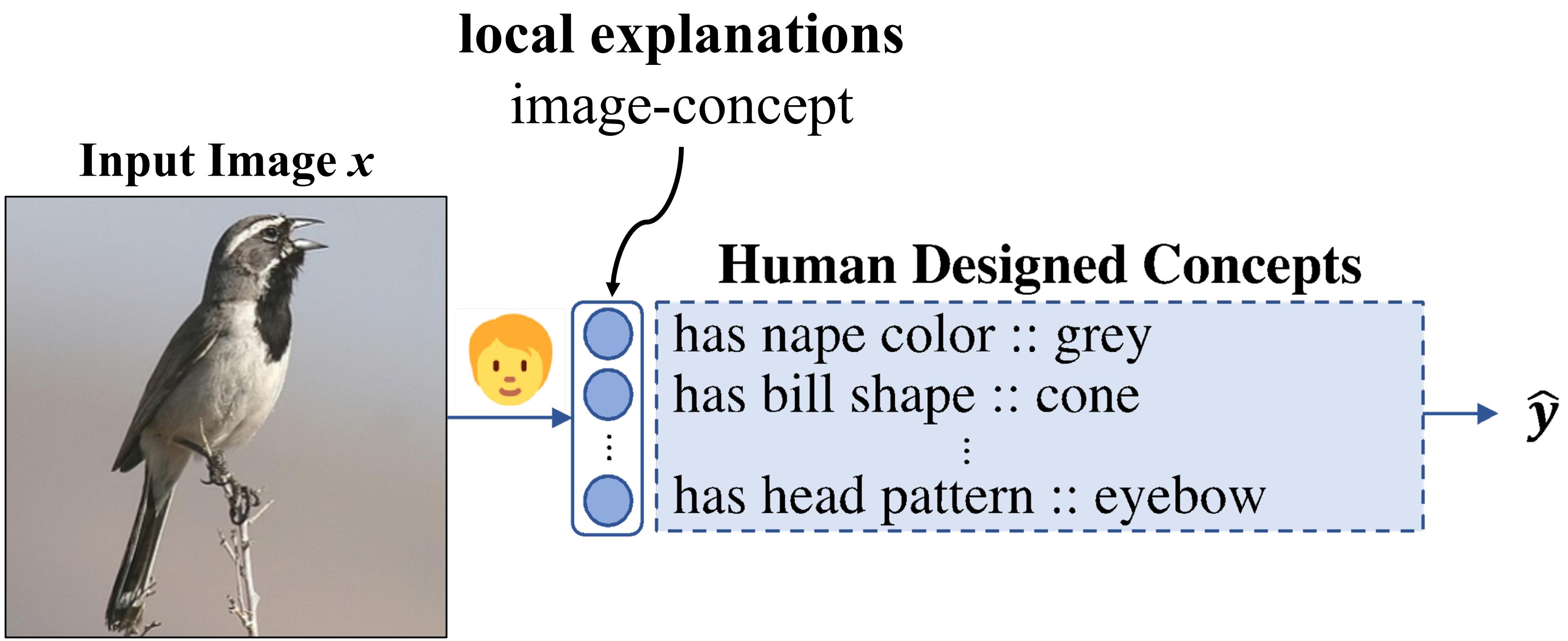
→ **label *y*** (black-throated sparrow)

Concept Bottleneck Models (CBMs)



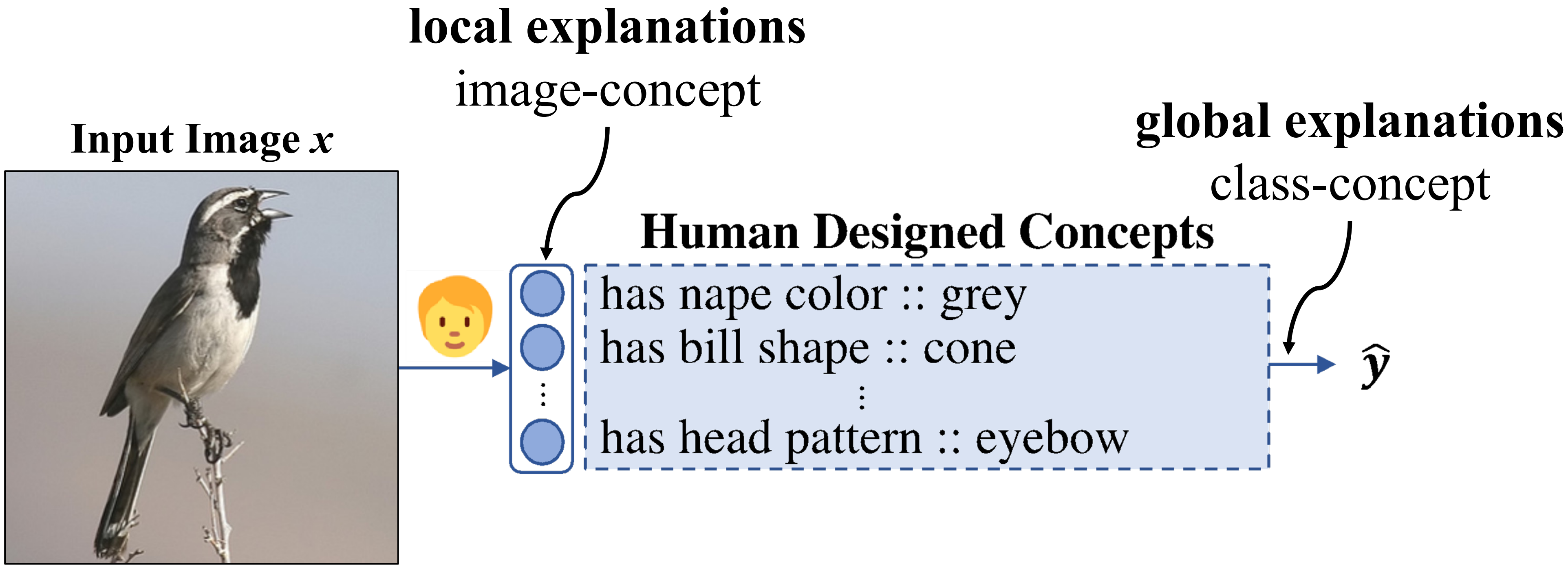
Koh et al. Concept Bottleneck Models. PMLR. 2020.

Concept Bottleneck Models (CBMs)

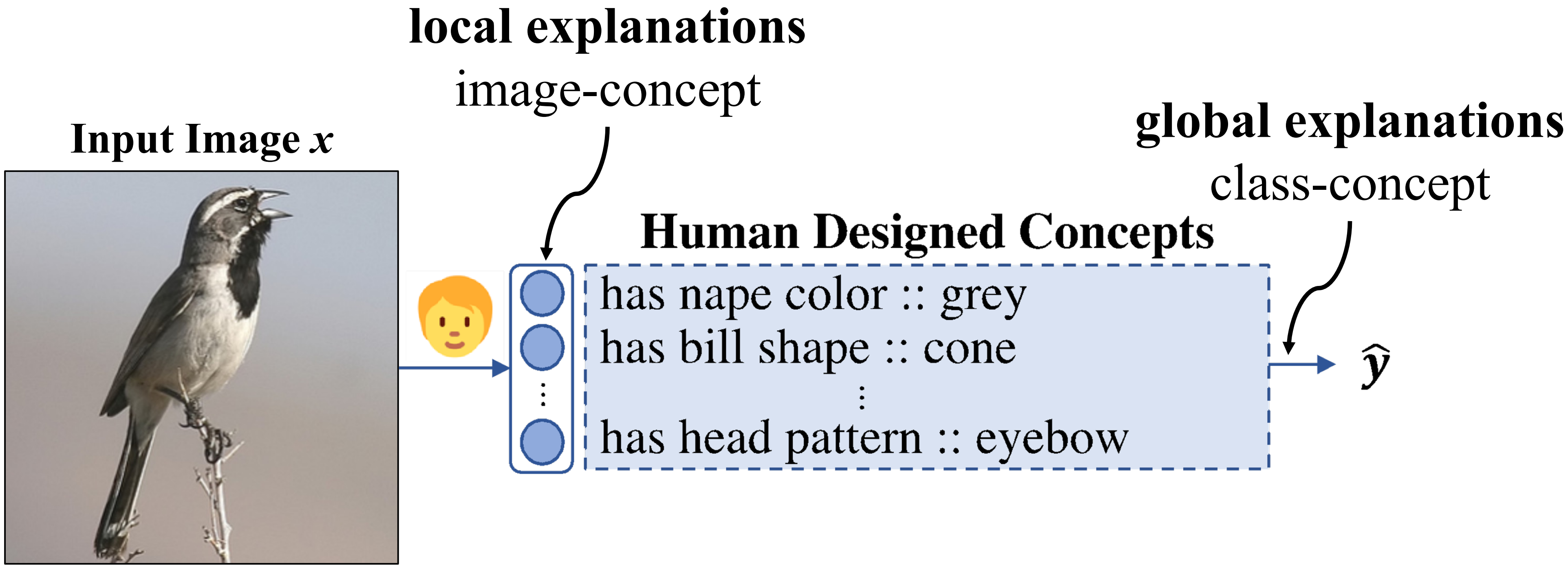


Koh et al. Concept Bottleneck Models. PMLR. 2020.

Concept Bottleneck Models (CBMs)



Concept Bottleneck Models (CBMs)



Challenges:

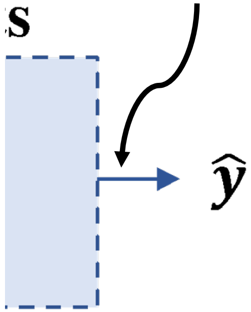
- **Scale**: requires human efforts in building concept bottlenecks.
- **Performance**: perform worse than black-box models.

Koh et al. Concept Bottleneck Models. PMLR. 2020.

Concept Bottleneck Models (CBMs)

Input Im

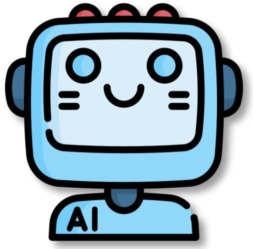
ɔbal explanations
class-concept
ɪs



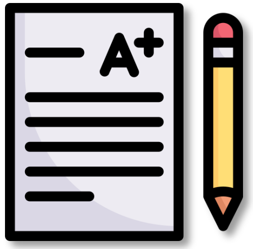
Challenge

- **Scale**: re ɔottlenecks.
- **Perforn** els.

Agenda of this talk

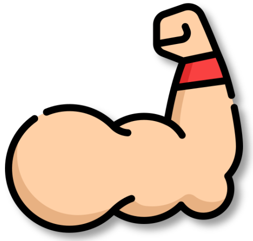


How to construct concept bottlenecks **without human effort**?



How to make interpretable models **performant** as black-box models?

LaBo (CVPR 23)



What other advantages can interpretable models give us? Answer: **Robustness**.

KnoBo
(In progress)

Language in a Bottle:
Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,
Chris Callison-Burch, Mark Yatskar

University of Pennsylvania

Penn Engineering
University *of* Pennsylvania

Leverage the world knowledge of LLMs

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Leverage the world knowledge of LLMs

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

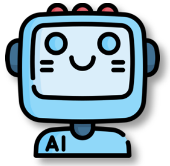
Leverage the world knowledge of LLMs

Describe what the *black-throated sparrow* looks like.

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Leverage the world knowledge of LLMs

Describe what the *black-throated sparrow* looks like.



GPT-3

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Leverage the world knowledge of LLMs

Describe what the *black-throated sparrow* looks like.

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

GPT-3

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Leverage the world knowledge of LLMs

Describe what the *black-throated sparrow* looks like.

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

GPT-3

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Leverage the world knowledge of LLMs

Describe what the *black-throated sparrow* looks like.

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

GPT-3

Describe the **shape/color** of the *black-throated sparrow*.

Brown et al. Language models are few-shot learners. NeurIPS. 2020.

Select the knowledge

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is <u>found in North America</u>. It is a <u>common bird in the western United States</u>. The black-throated sparrow is <u>a member of the sparrow family.</u>

Select the knowledge

too general

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is <u>found in North America</u>. It is a <u>common bird in the western United States</u>. The black-throated sparrow is <u>a member of the sparrow family.</u>

Select the knowledge

too general

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

not visual

Select the knowledge

too general

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

not visual

Submodular Optimization

Visual
Discriminative
Diverse

Select the knowledge

too general

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

not visual

Submodular Optimization

Visual
Discriminative
Diverse

Select the knowledge

too general

The black-throated sparrow is a **small bird** with a **black head and throat**. It has a **white body with brown streaks on its back**. Its **wings are brown with white stripes**. The black-throated sparrow has **a long, thin beak**. It **has two long, thin legs**. The black-throated sparrow has a **long, thin tail**. It is about **5 inches long**. The black-throated sparrow is found in North America. It is a common bird in the western United States. The black-throated sparrow is a member of the sparrow family.

not visual

Submodular Optimization

Visual
Discriminative
Diverse

black head and throat

long, thin tail

wings are brown with white stripes

White body with brown streaks

Ground Concepts using CLIP

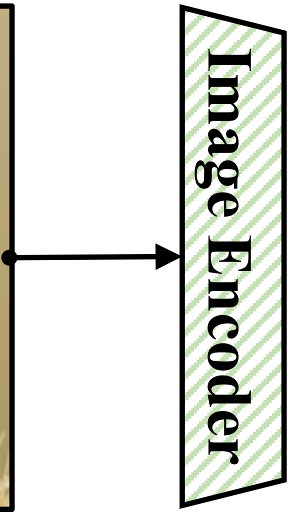
black head and throat

long, thin tail

wings are brown with white stripes

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR. 2021.

Ground Concepts using CLIP



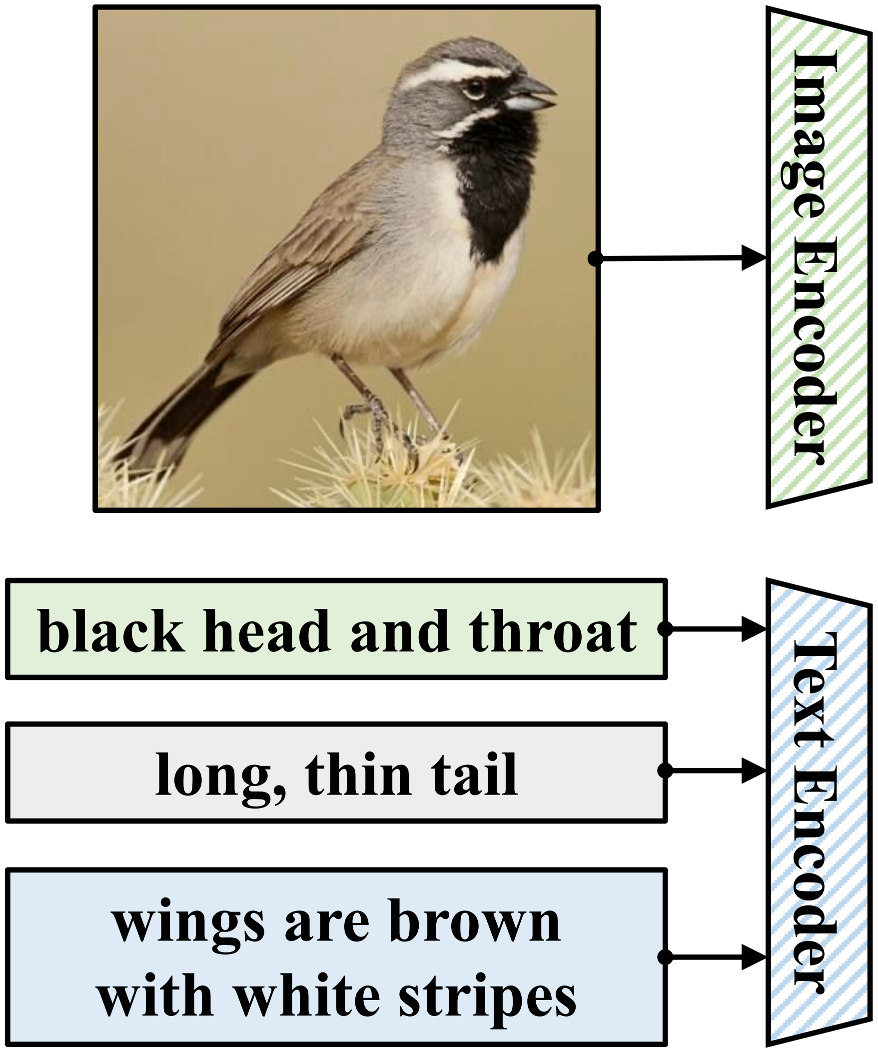
black head and throat

long, thin tail

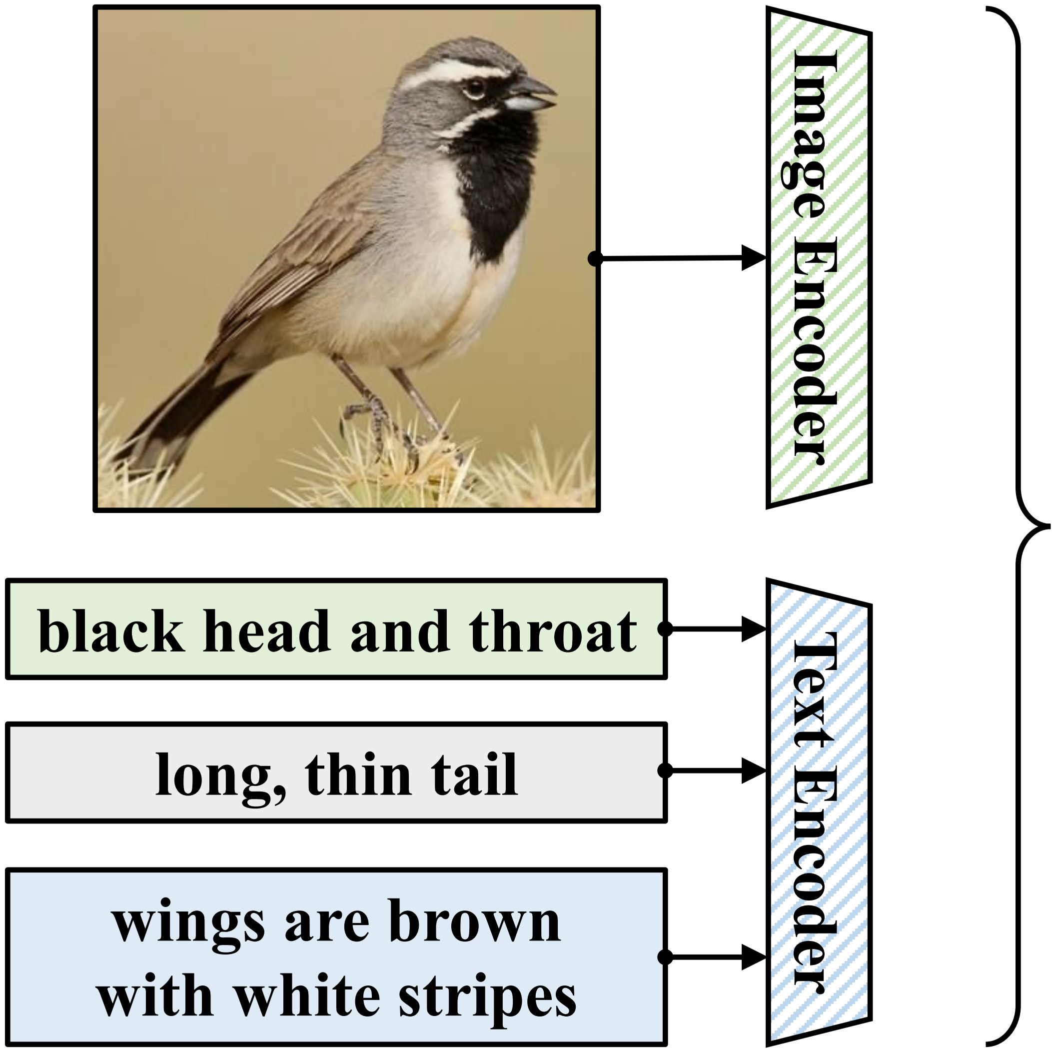
wings are brown
with white stripes

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR. 2021.

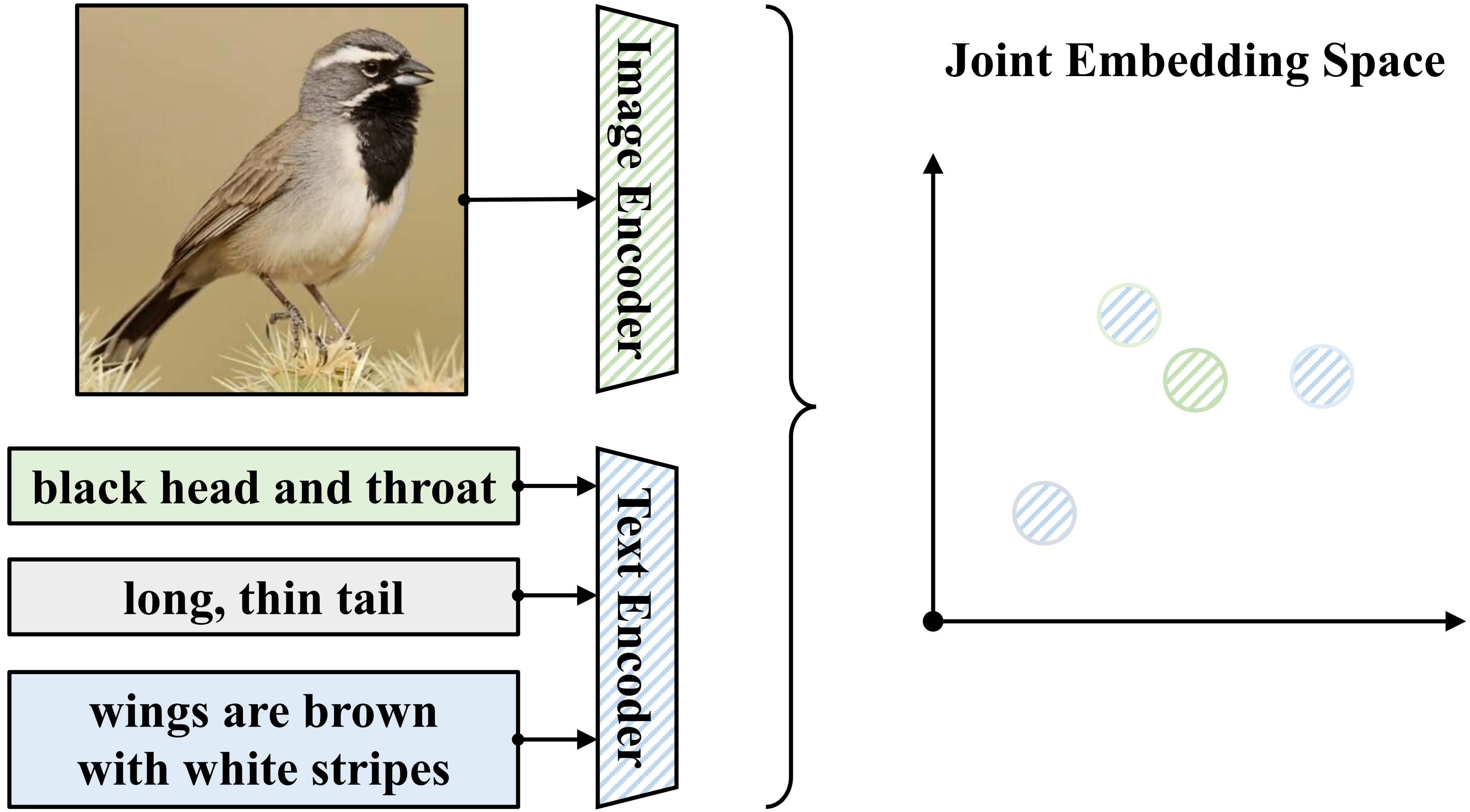
Ground Concepts using CLIP



Ground Concepts using CLIP

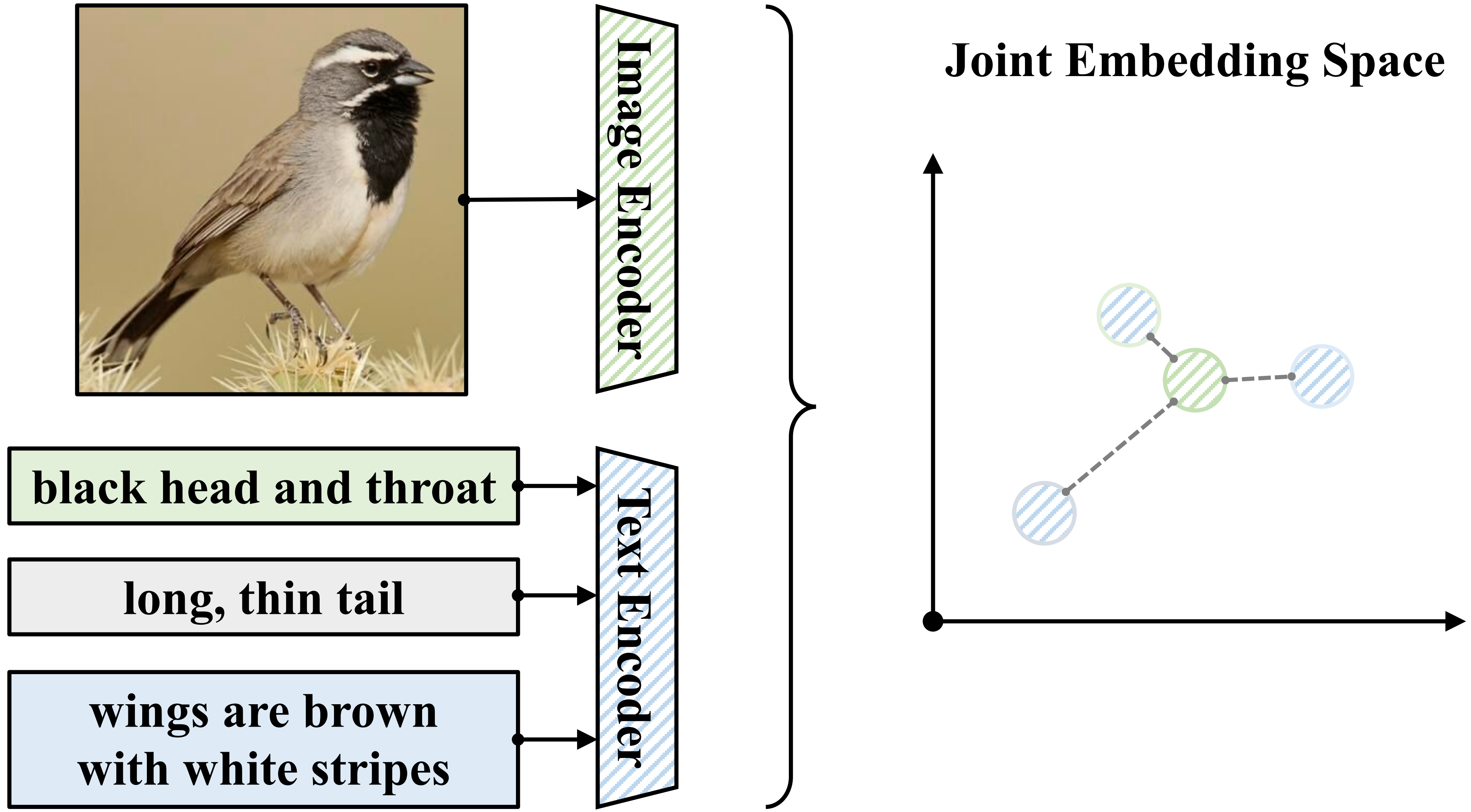


Ground Concepts using CLIP



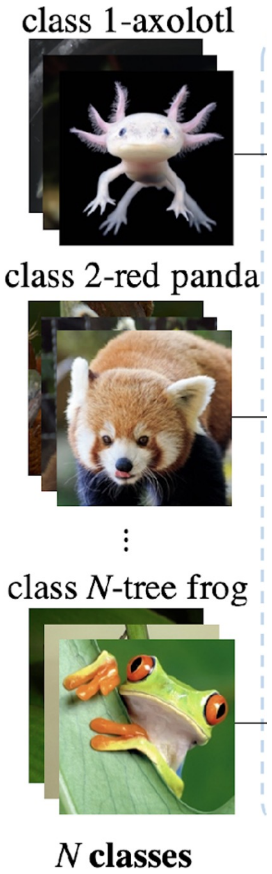
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR. 2021.

Ground Concepts using CLIP



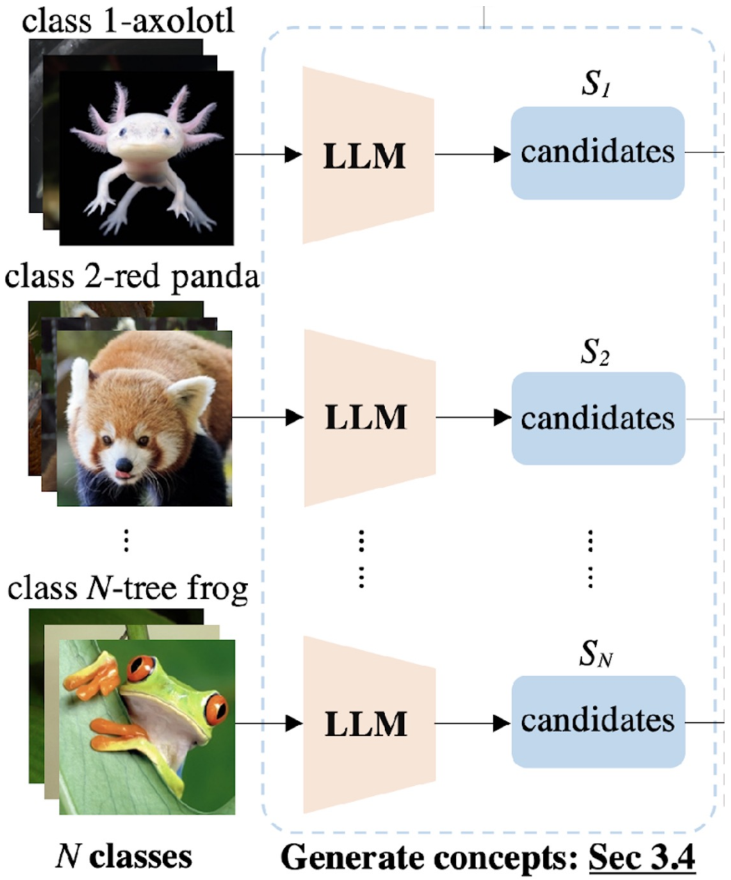
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR. 2021.

Prompt LLM to generate candidate concepts



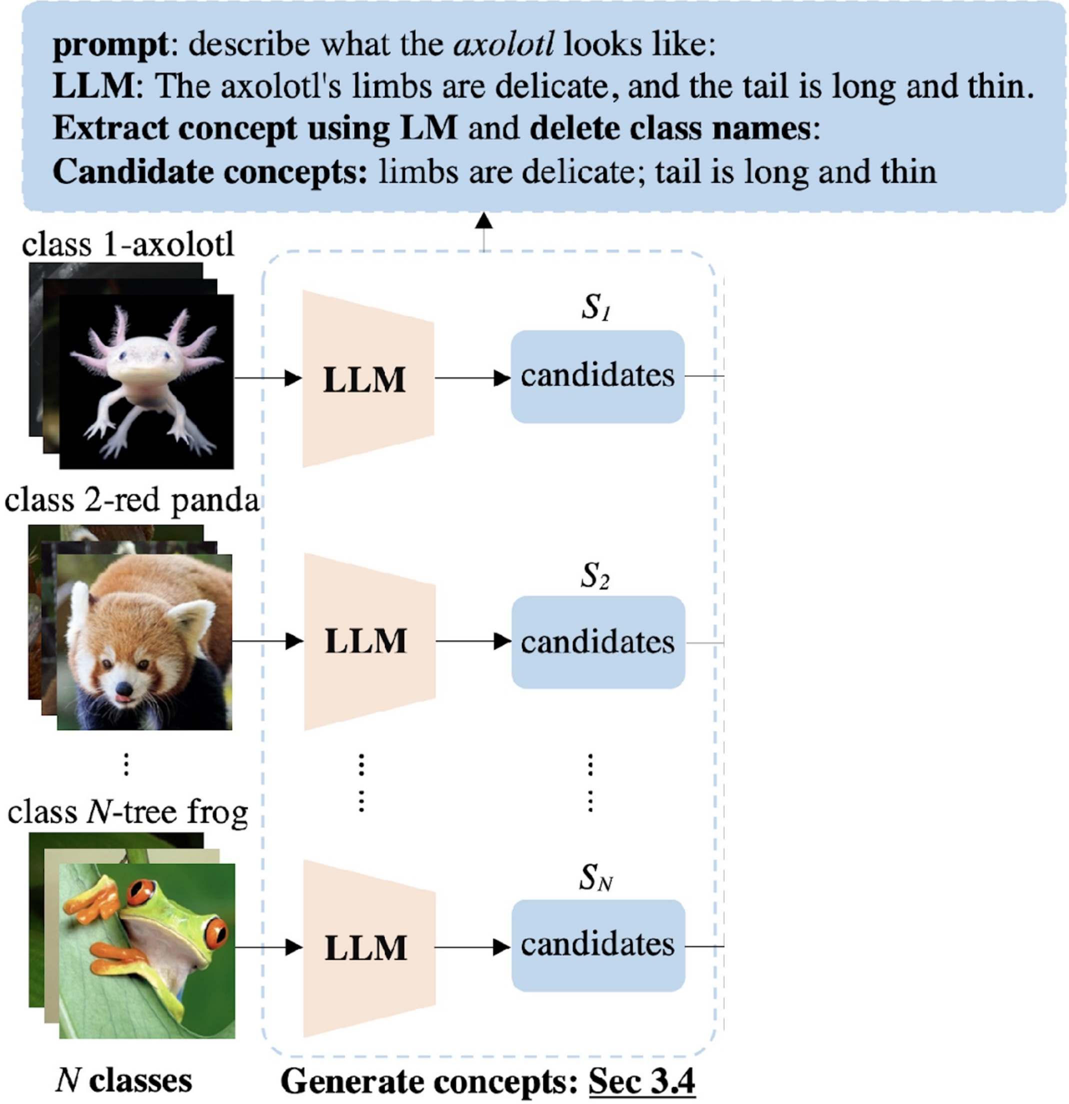
[1] Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020.

Prompt LLM to generate candidate concepts



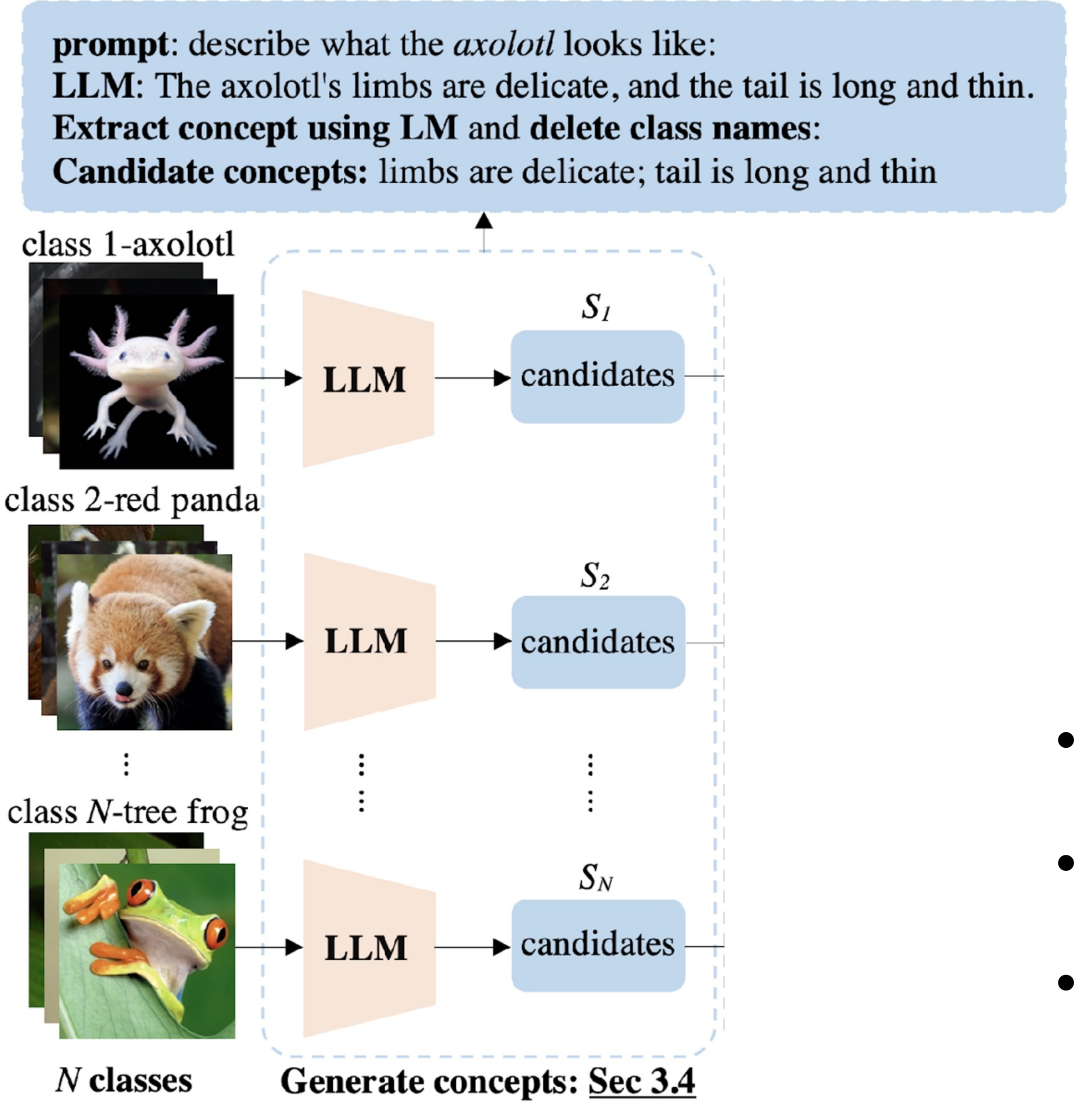
[1] Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020.

Prompt LLM to generate candidate concepts



[1] Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020.

Prompt LLM to generate candidate concepts

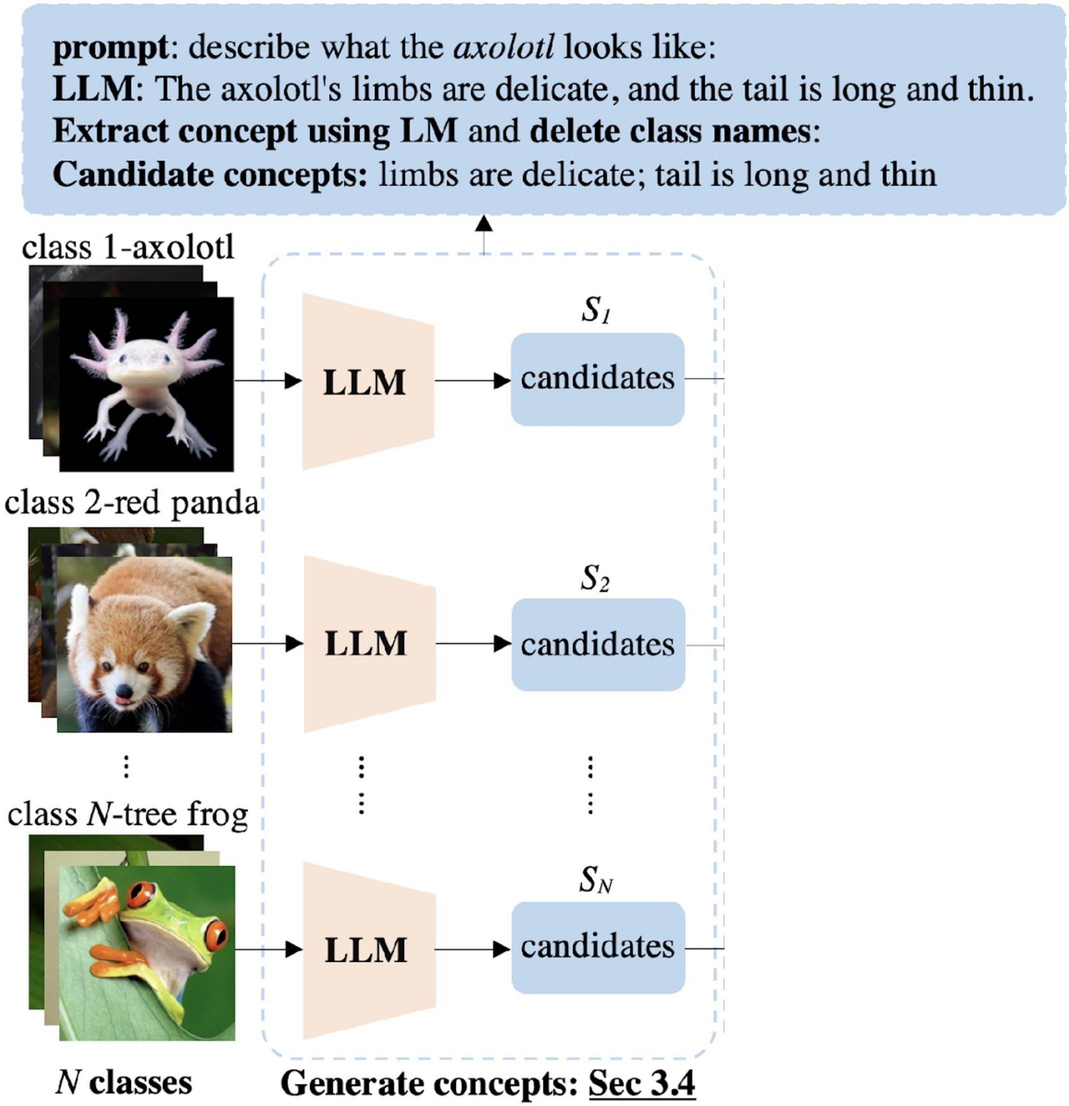


General Prompt Template
1. describe what the `[CLASS NAME]` looks like:
2. describe the appearance of the `[CLASS NAME]`:
3. describe the color of the `[CLASS NAME]`:
4. describe the pattern of the `[CLASS NAME]`:
5. describe the shape of the `[CLASS NAME]`:

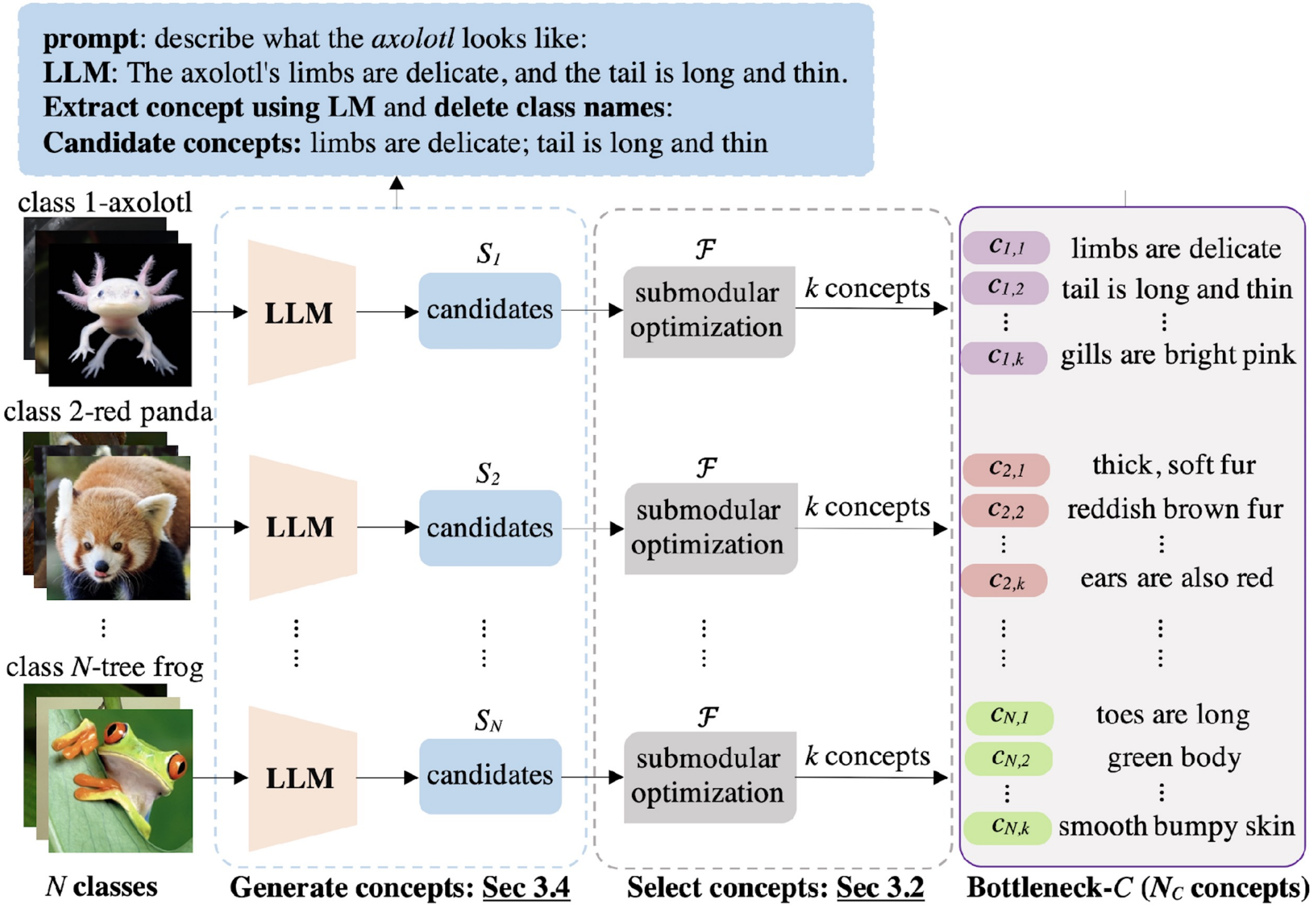
- Obtain 500 sentences for each class.
- Extract concepts from sentences using T5 [1].
- String match to identify and remove class name tokens in each concept.

[1] Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020.

Submodular Concept Selection



Submodular Concept Selection



Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

candidate S_y
selected C_y

Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

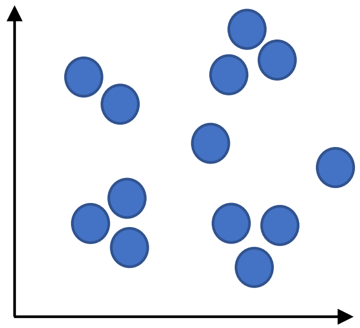
$$\mathcal{F}(C_y) = \underbrace{\alpha \cdot \sum_{c \in C_y} D(c)}_{\text{discriminability}} + \underbrace{\beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$

candidate S_y
selected C_y

Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

$$\mathcal{F}(C_y) = \underbrace{\alpha \cdot \sum_{c \in C_y} D(c)}_{\text{discriminability}} + \underbrace{\beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$



candidate S_y
selected C_y

Submodular Concept Selection

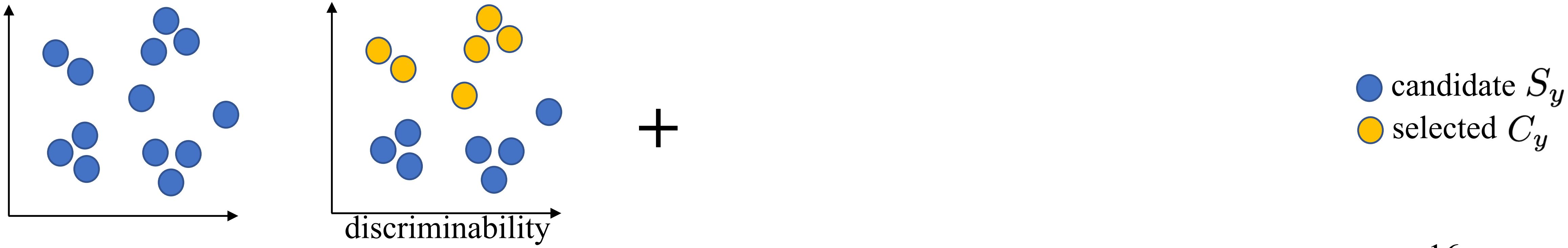
- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

$$\mathcal{F}(C_y) = \underbrace{\alpha \cdot \sum_{c \in C_y} D(c)}_{\text{discriminability}} + \underbrace{\beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$

Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

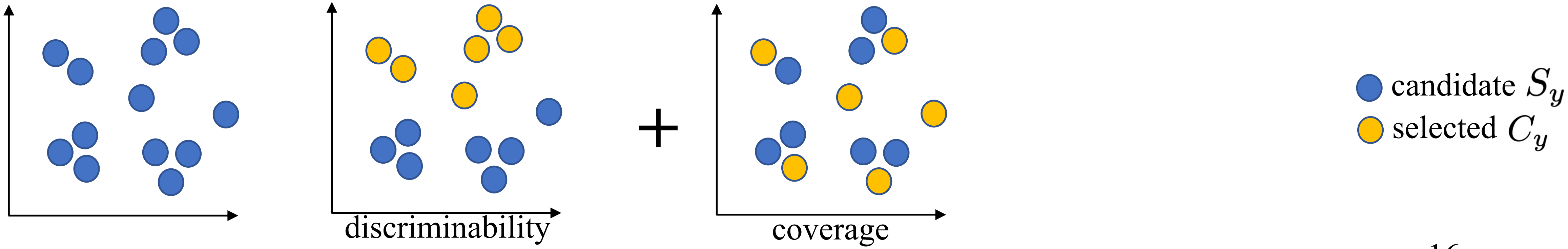
$$\mathcal{F}(C_y) = \alpha \cdot \underbrace{\sum_{c \in C_y} D(c)}_{\text{discriminability}} + \beta \cdot \underbrace{\sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$



Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

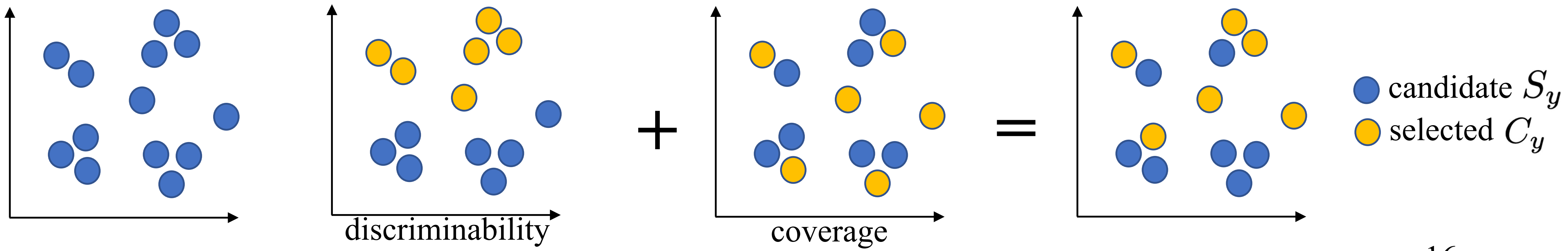
$$\mathcal{F}(C_y) = \underbrace{\alpha \cdot \sum_{c \in C_y} D(c)}_{\text{discriminability}} + \underbrace{\beta \cdot \sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$



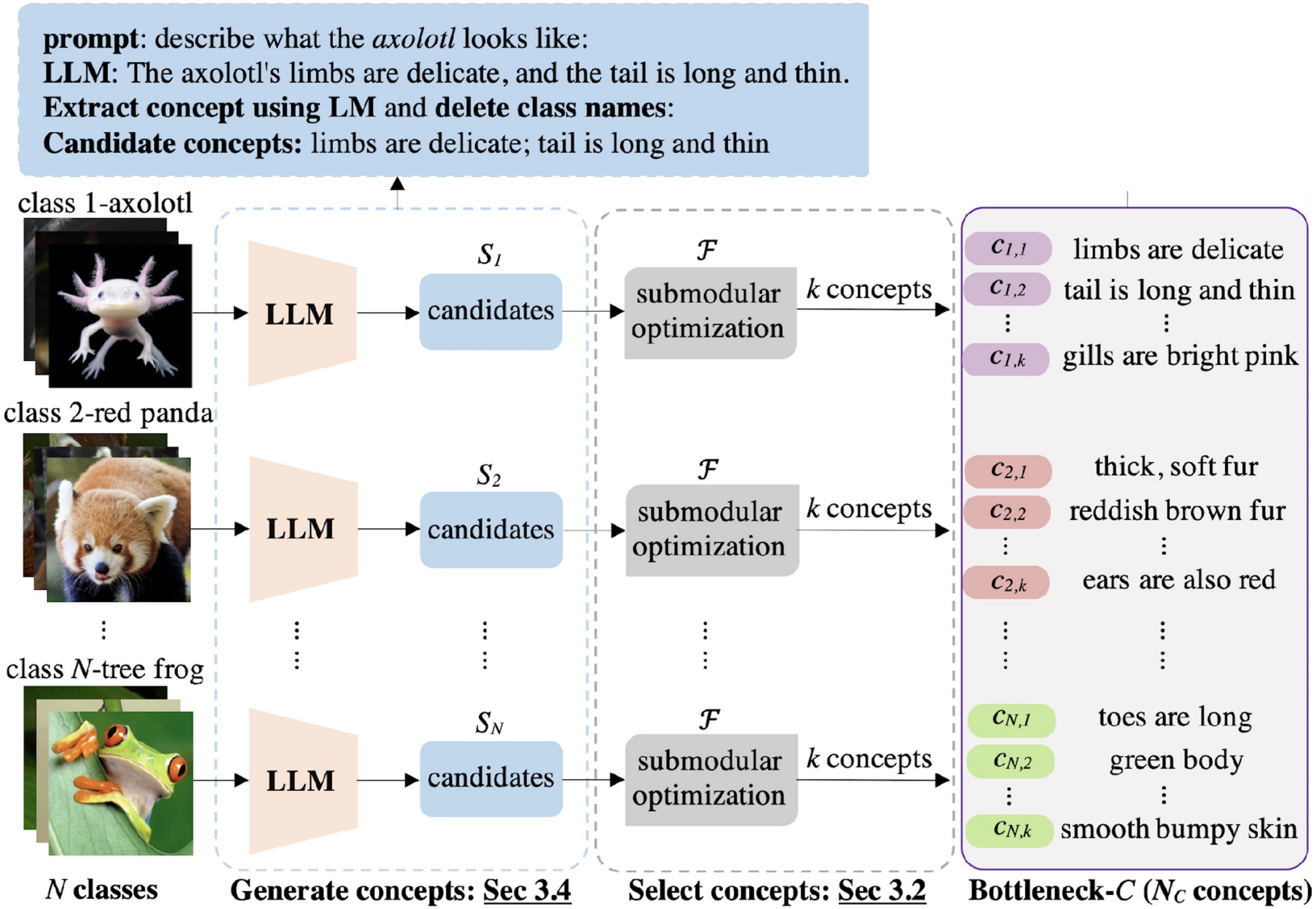
Submodular Concept Selection

- Given a superset of concepts S_y for a class y.
- Select a subset C_y for the bottleneck which are **discriminative** and **diverse**.

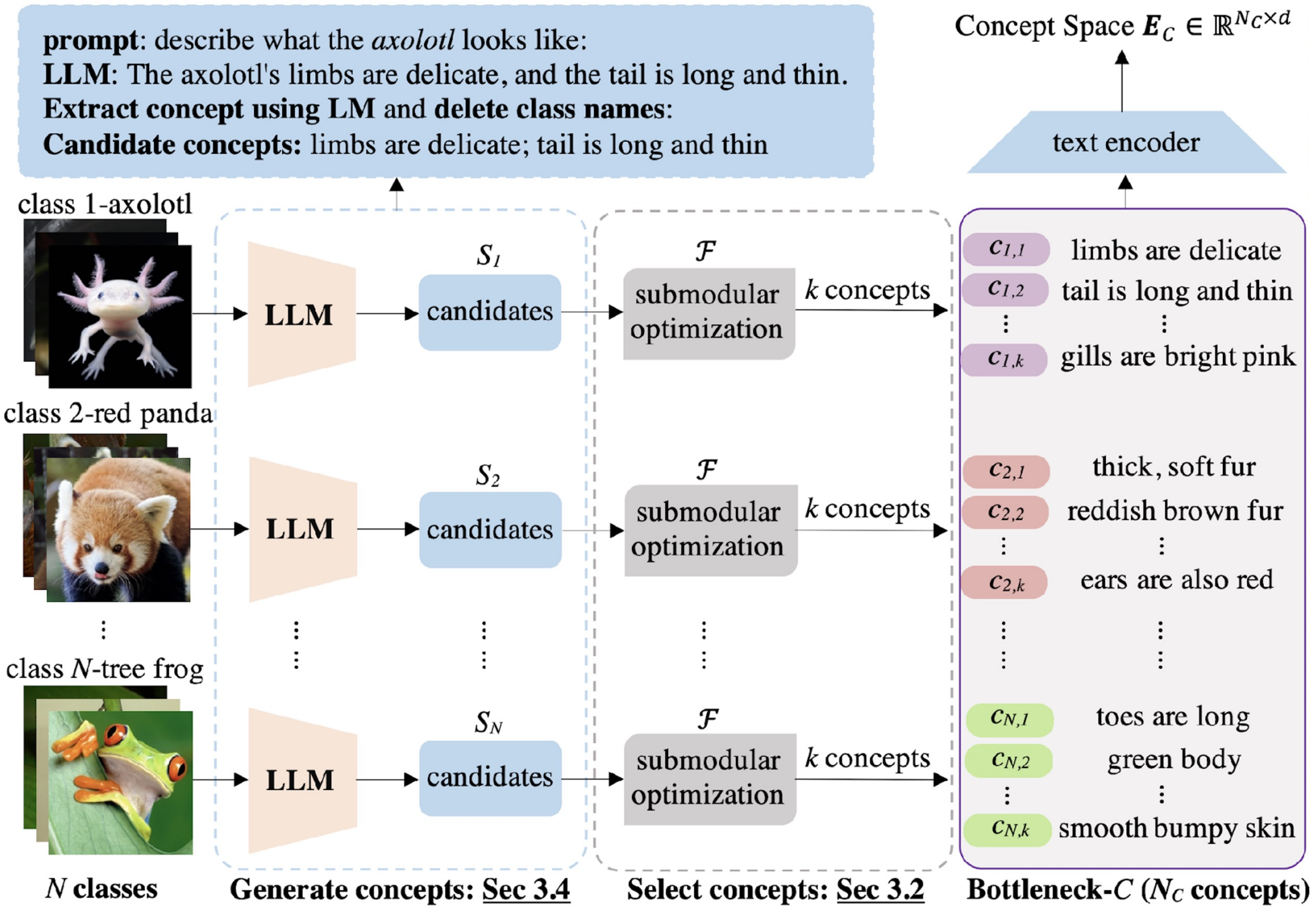
$$\mathcal{F}(C_y) = \alpha \cdot \underbrace{\sum_{c \in C_y} D(c)}_{\text{discriminability}} + \beta \cdot \underbrace{\sum_{c_1 \in S_y} \max_{c_2 \in C_y} \phi(c_1, c_2)}_{\text{coverage}}$$



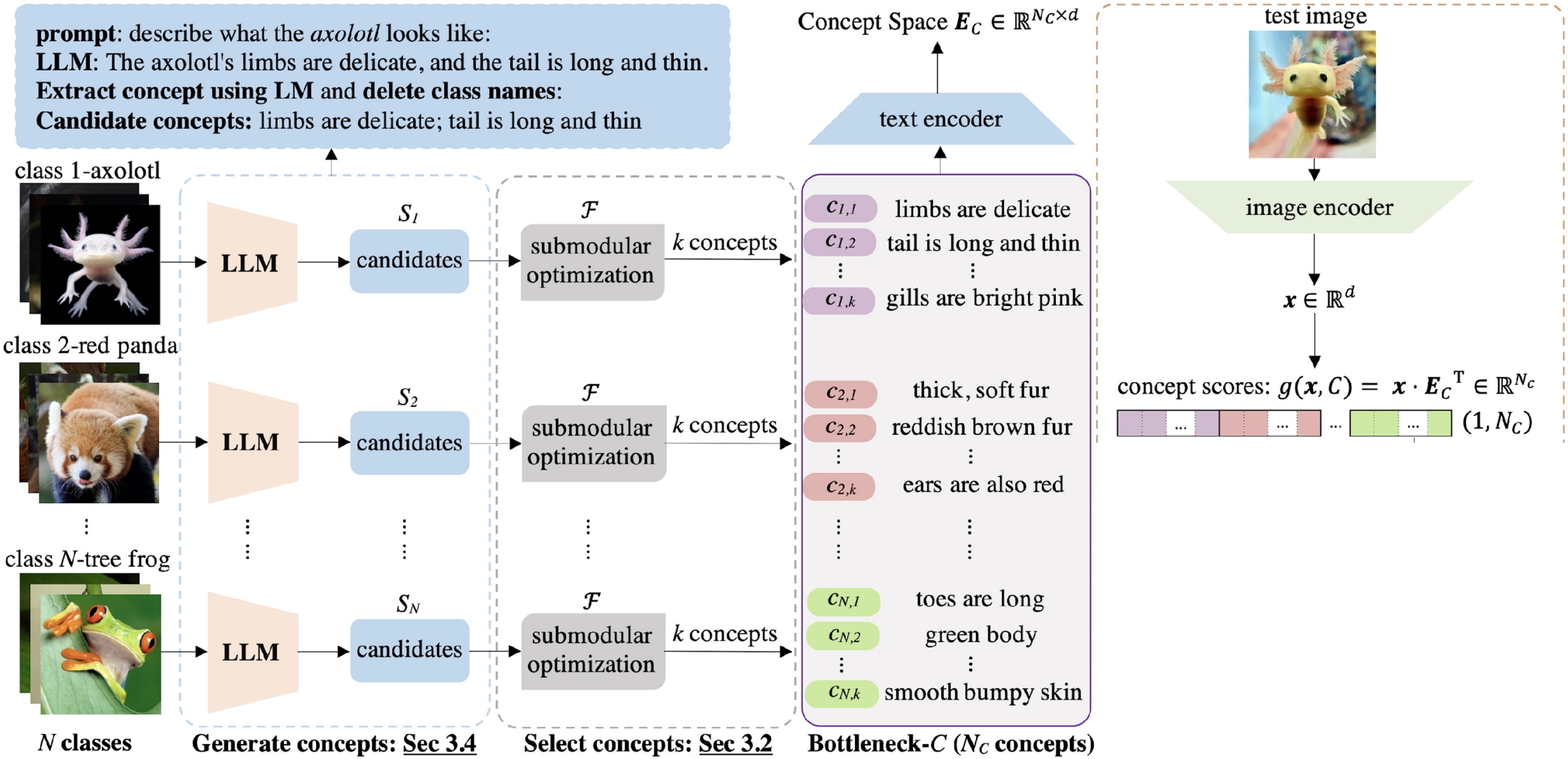
Compute Concept Scores



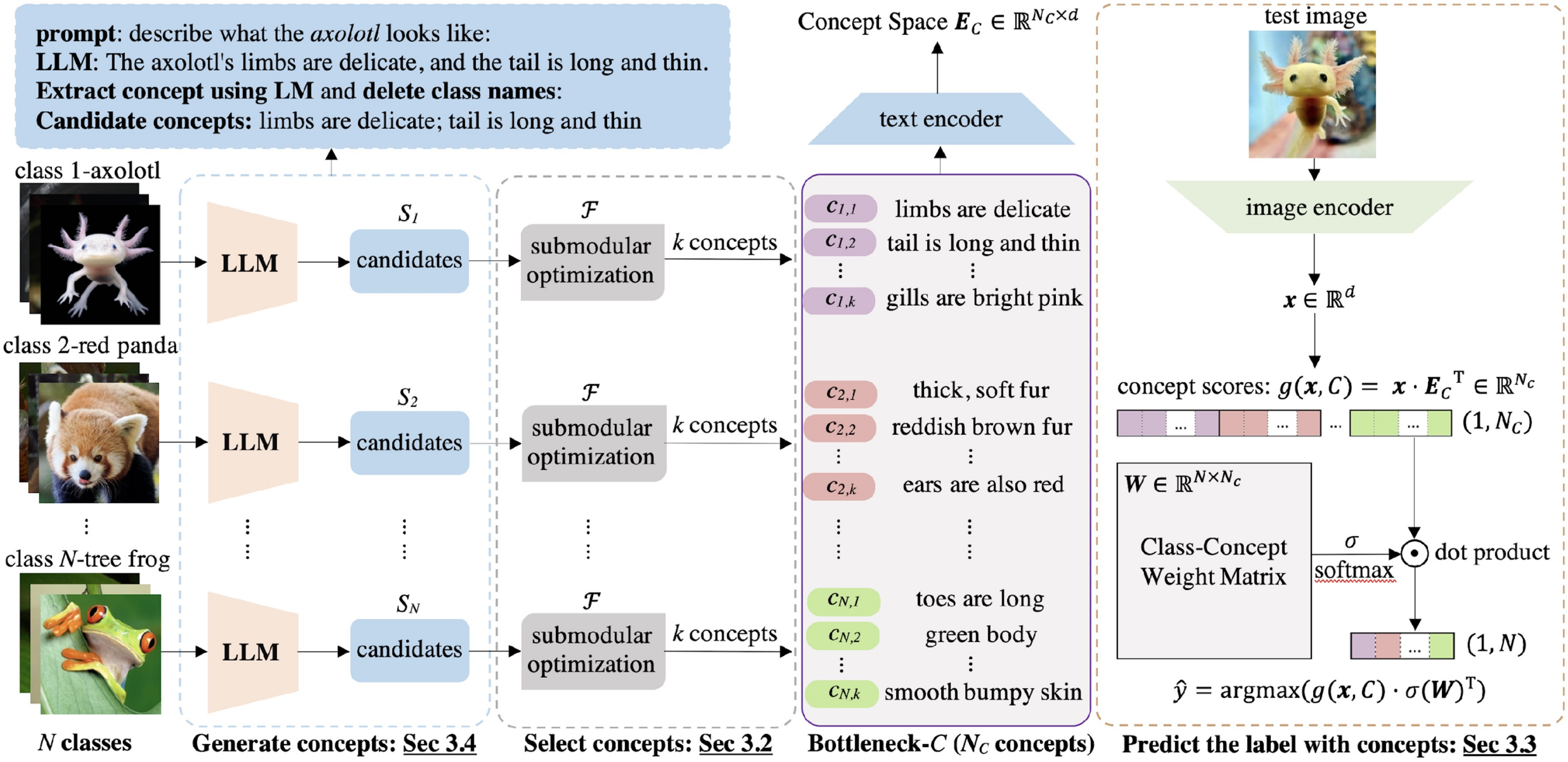
Compute Concept Scores



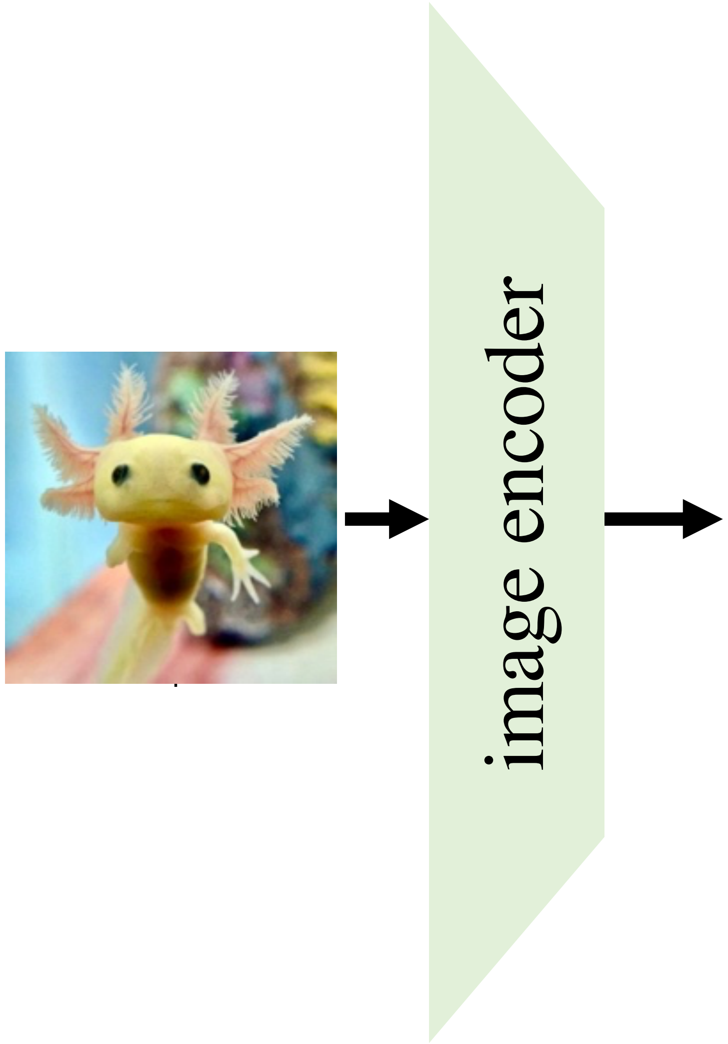
Compute Concept Scores



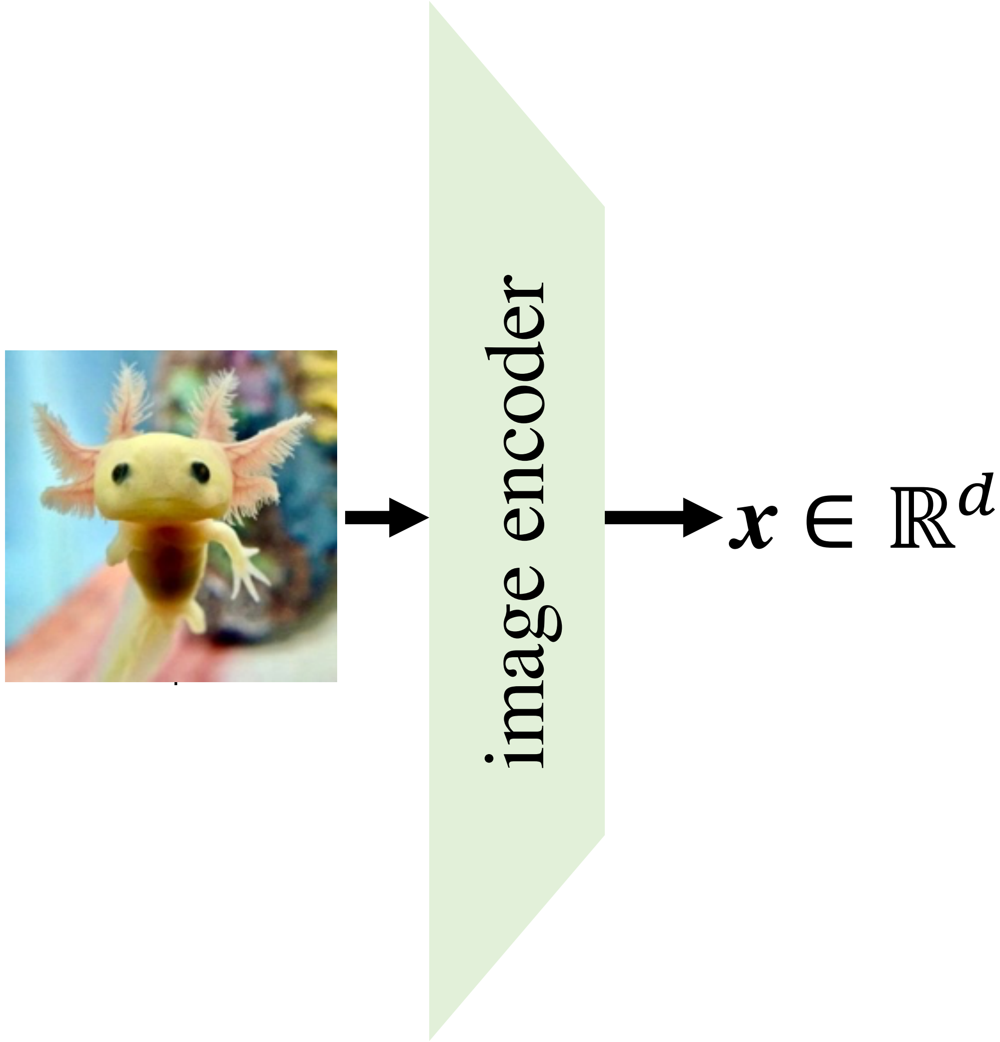
Compute Concept Scores



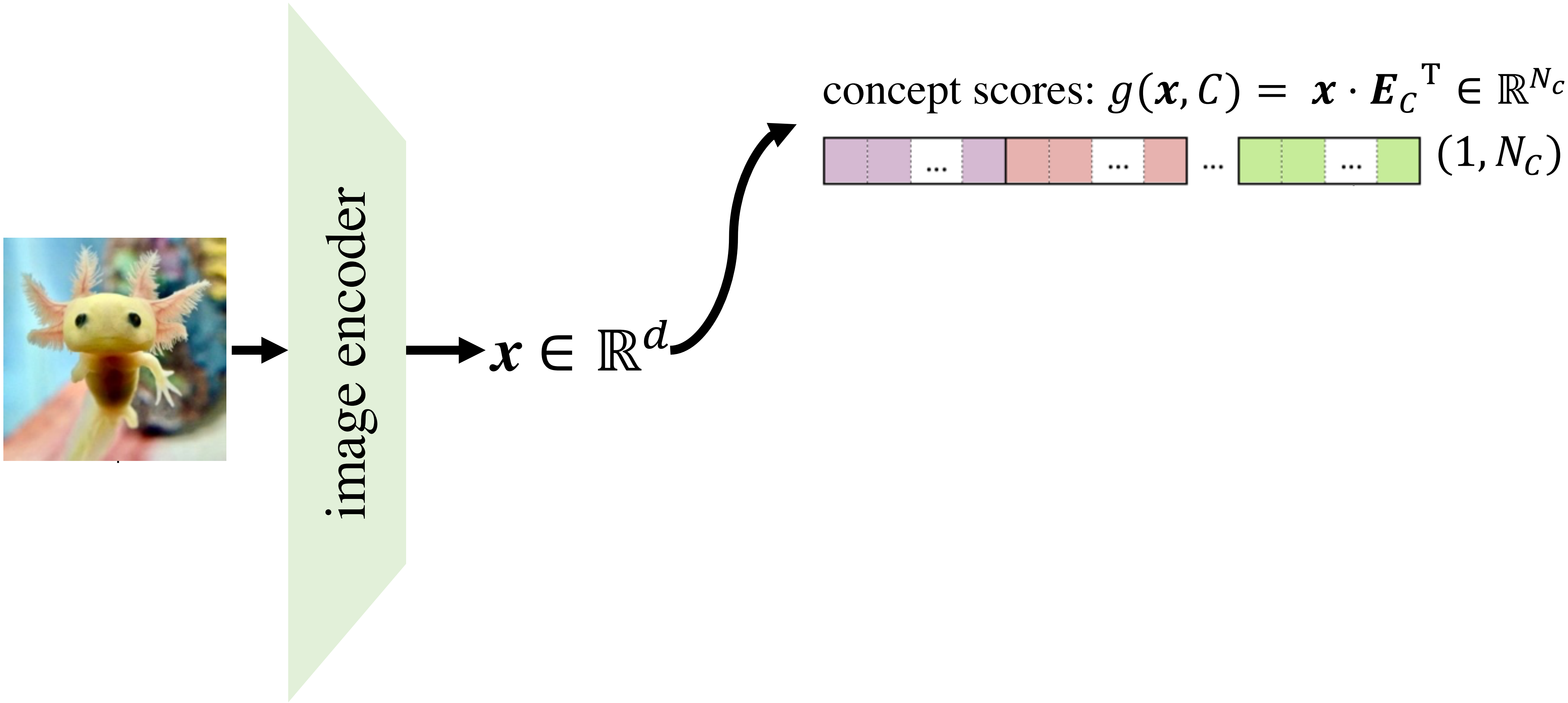
Concept scores and label prediction



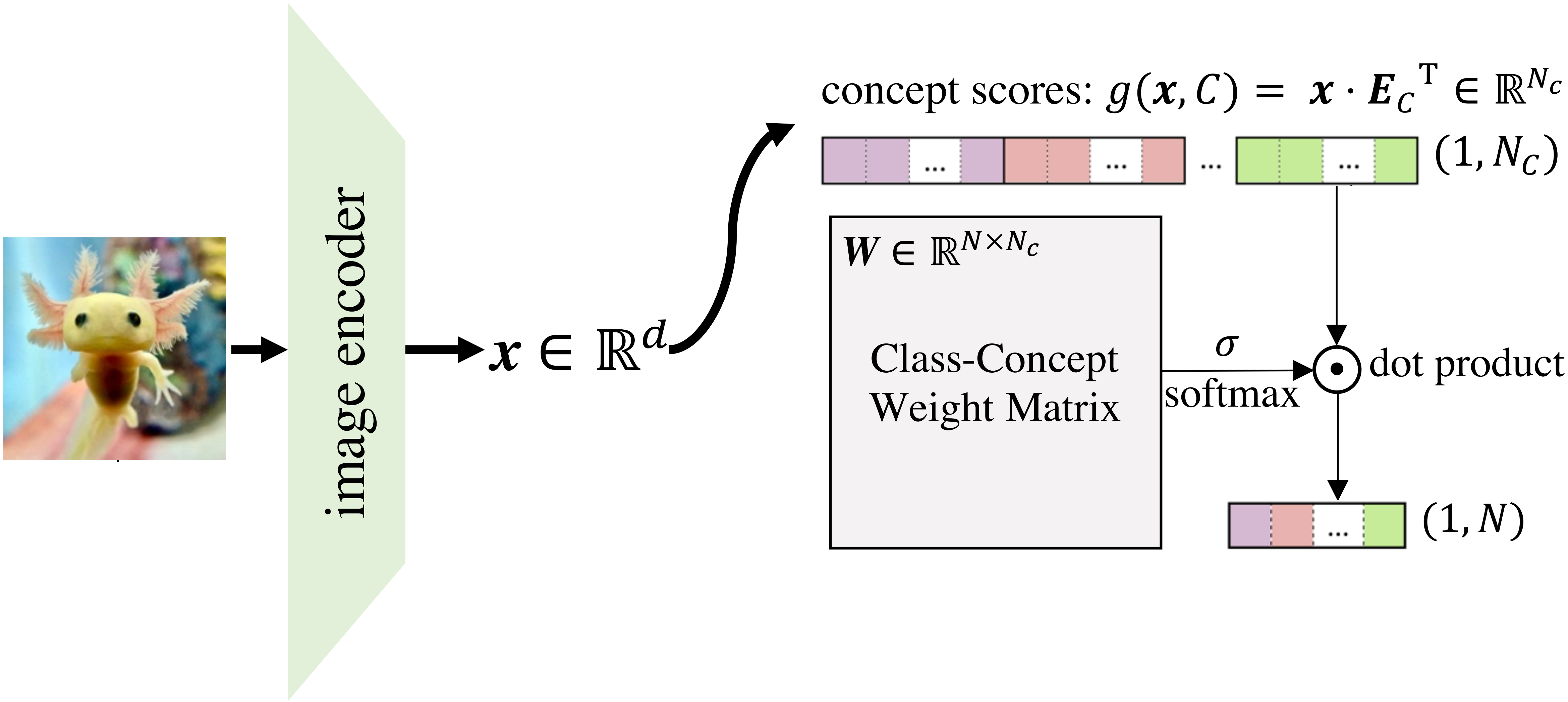
Concept scores and label prediction



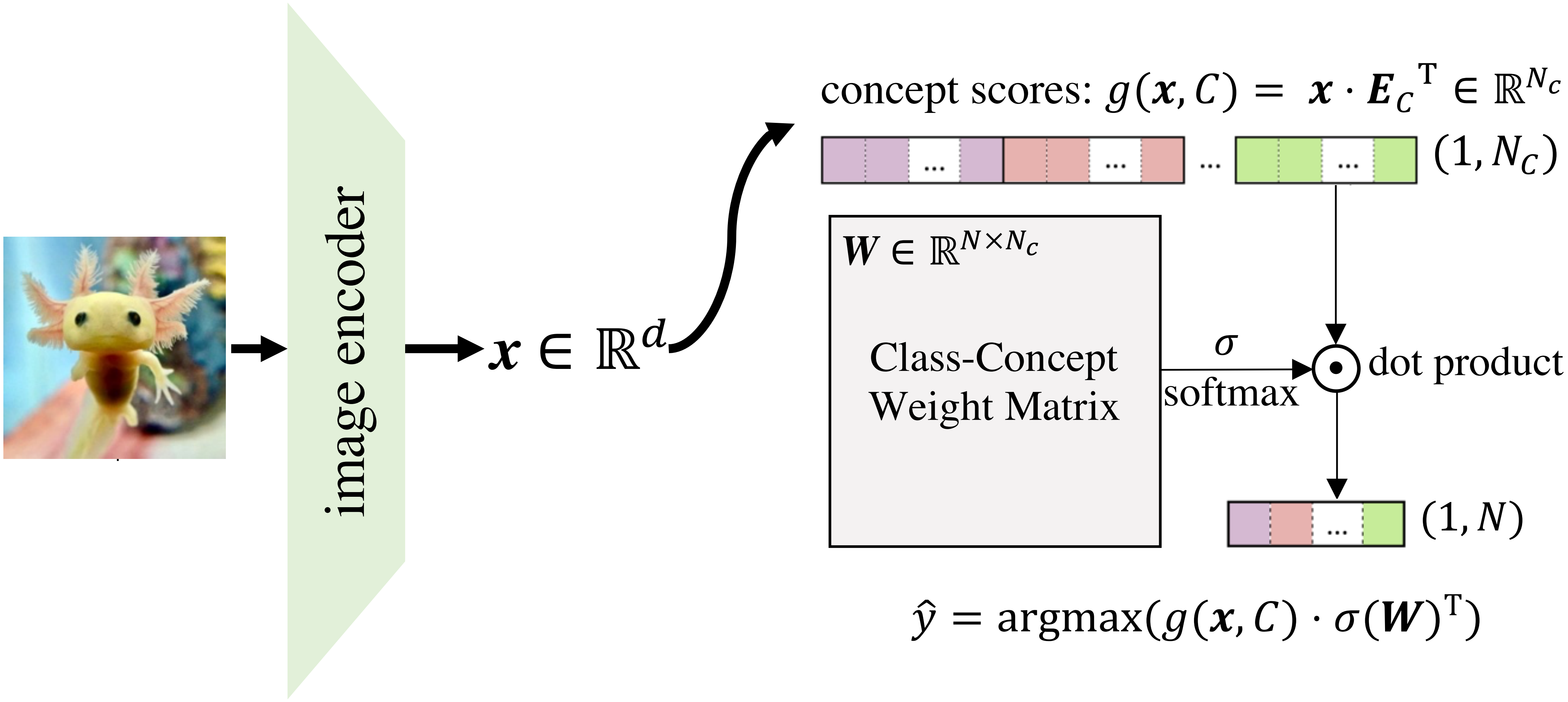
Concept scores and label prediction



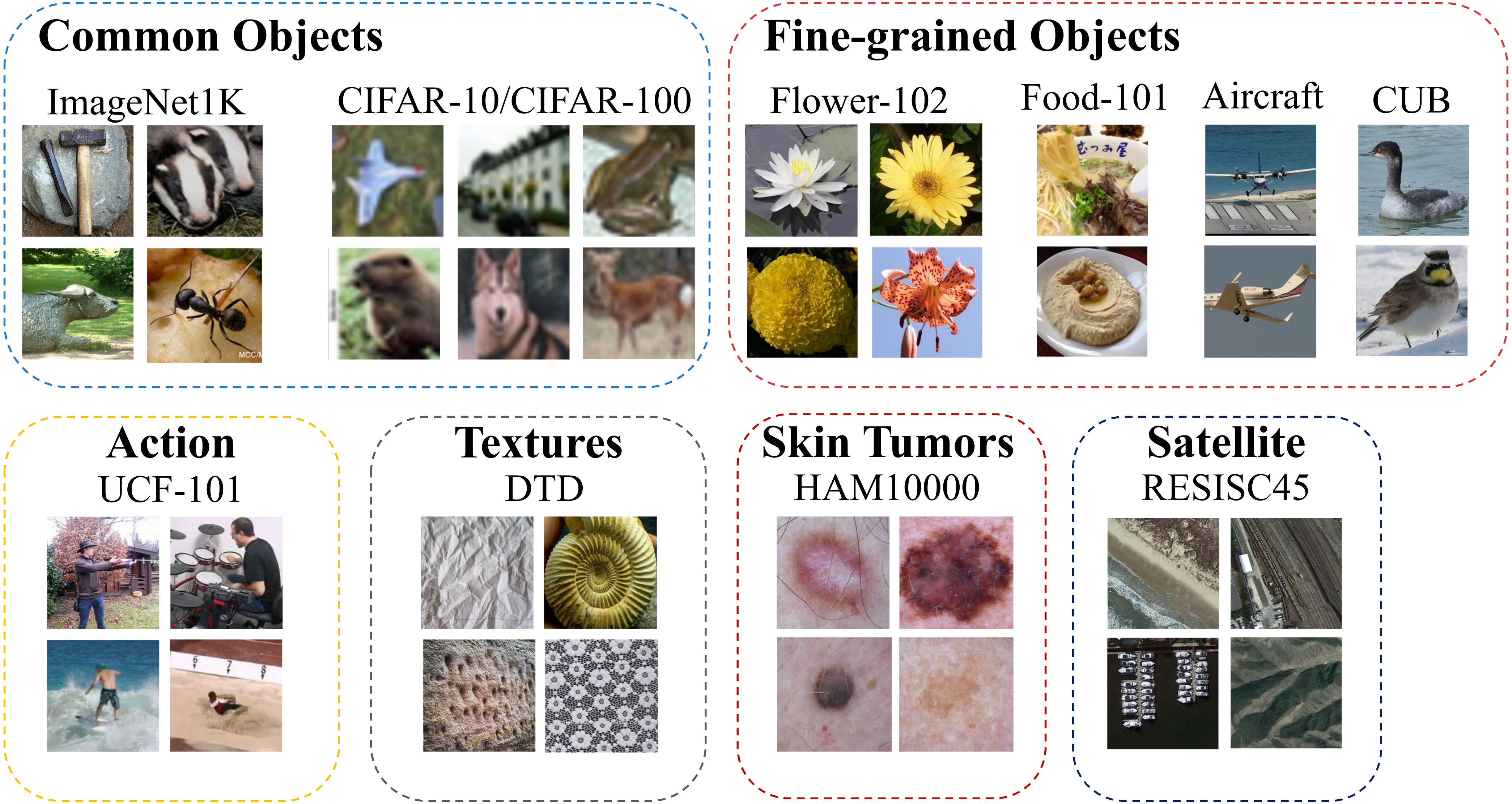
Concept scores and label prediction



Concept scores and label prediction



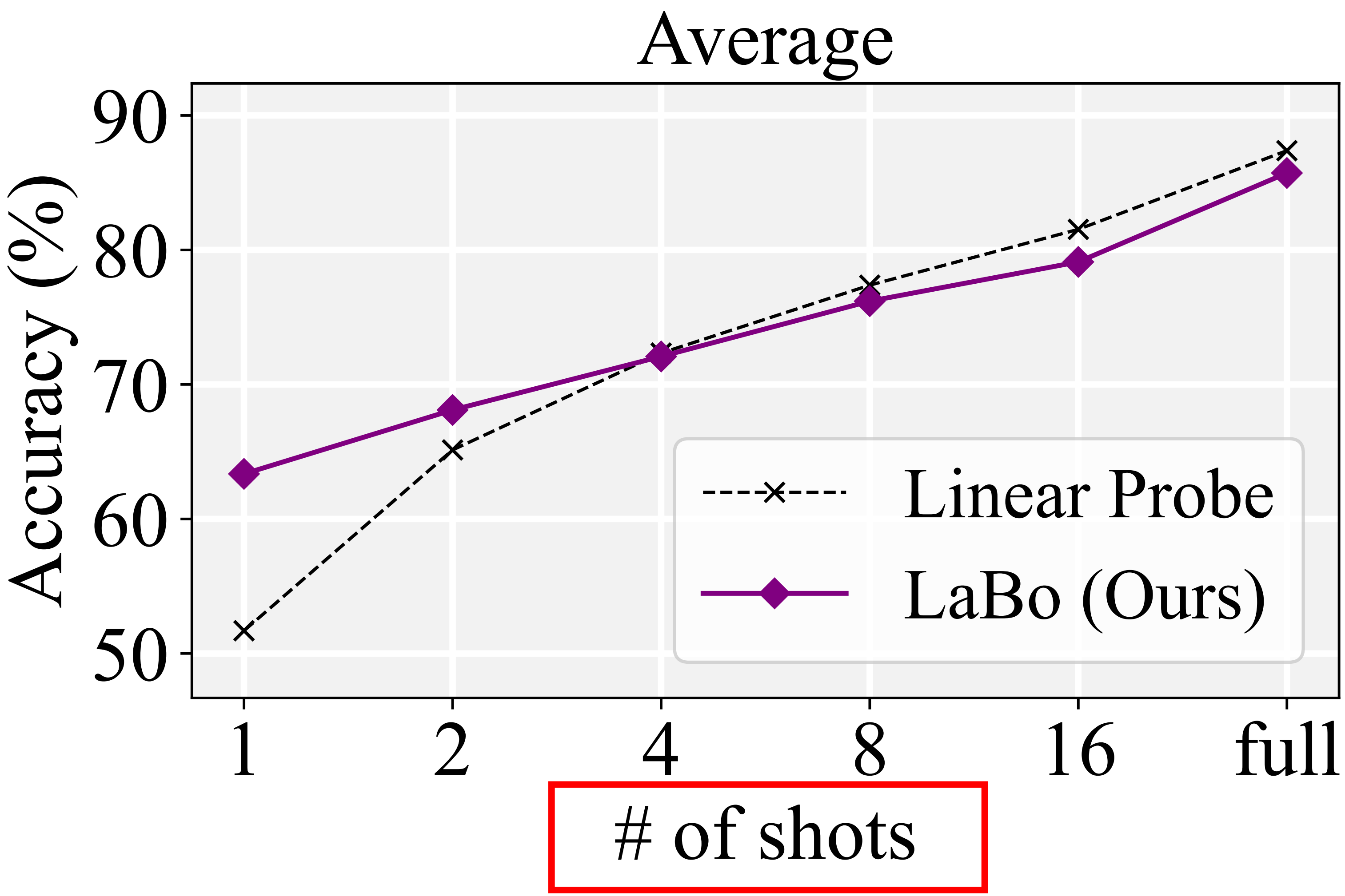
Datasets


Common Objects
ImageNet1K
CIFAR-10/CIFAR-100
Fine-grained Objects
Flower-102
Food-101
Aircraft
CUB
Action
UCF-101
Textures
DTD
Skin Tumors
HAM10000
Satellite
RESISC45

Experimental Setup

- **Baselines**:
  - Linear Probe: logistic regression on the image features.
  - PCBM: Post-hoc CBM (Yuksekgonul et al., 2022)
    - Ensemble CBM prediction with end-to-end prediction.
  - ComDL: Compositional Derivation Learning (Yun et al., 2022)
    - Human designed concepts.
    - Linear layer over CLIP similarity scores.
- **Few-shot/Fully-supervised.**
- **Metric**: accuracy.

Comparison to Black-box Model



# of training images per class

Comparison to Blackbox Model

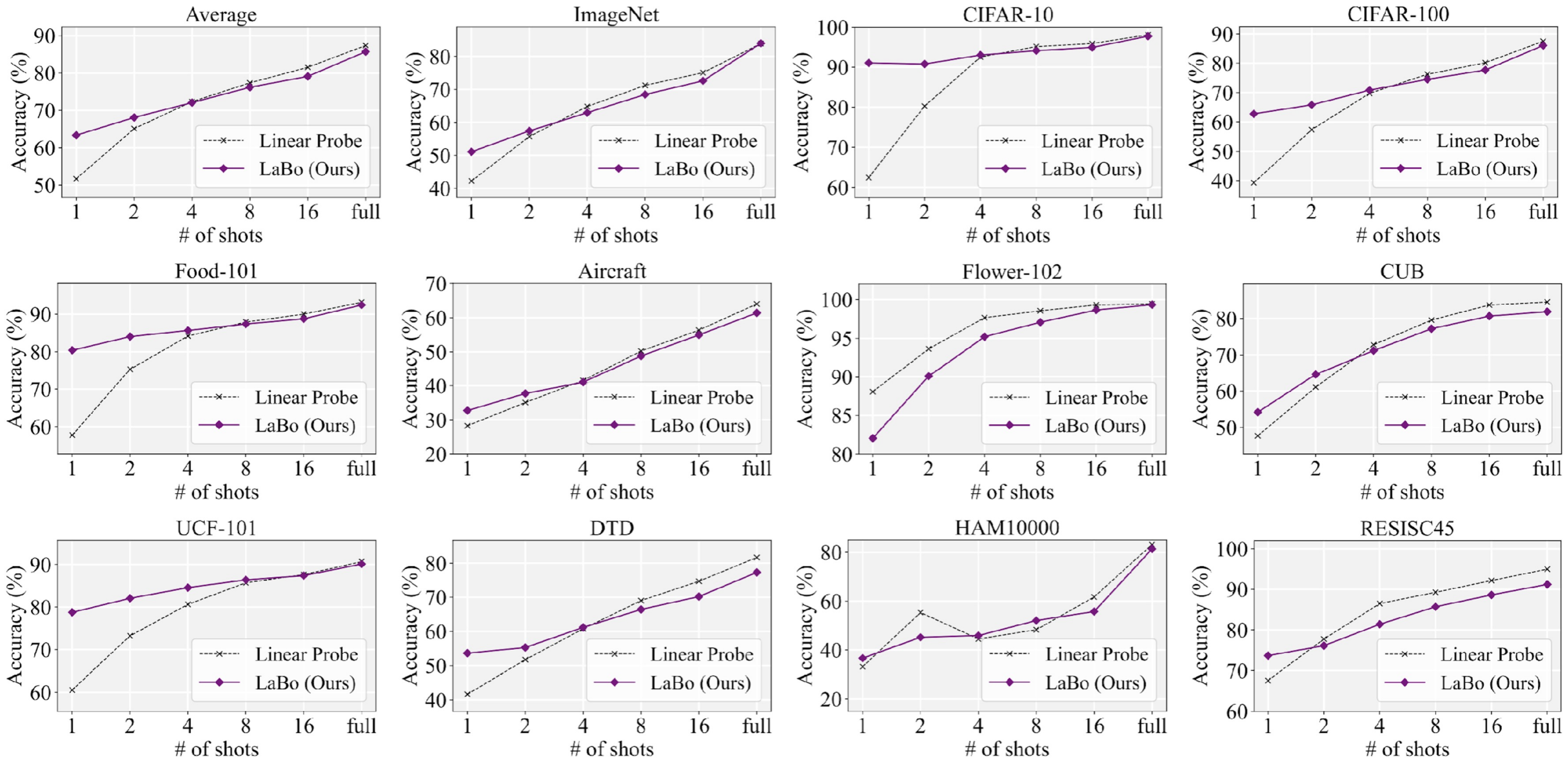


Figure 3. Test accuracy (%) comparison between LaBo and Linear Probe on 11 datasets. The x-axis represents the number of labeled images.

Comparison to Blackbox Model

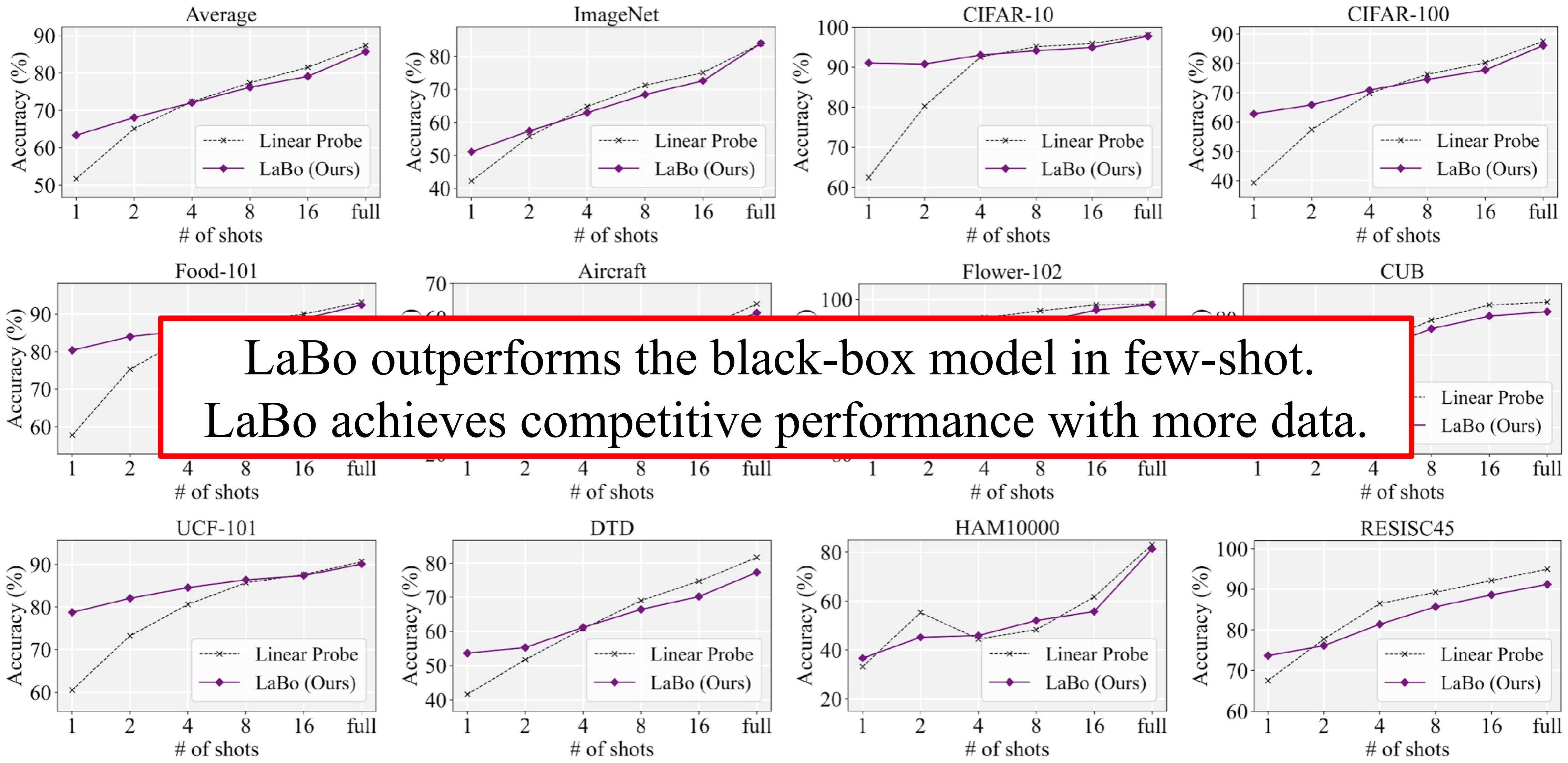


Figure 3. Test accuracy (%) comparison between LaBo and Linear Probe on 11 datasets. The x-axis represents the number of labeled images.

Compare with Previous CBM

Method	w/ end-to-end	CIFAR-10	CIFAR-100
PCBM [66]	✗	84.5	56.0
LaBo (Ours)	✗	**87.9**	**69.1**
PCBM-h [66]	✓	87.6	69.9
Linear Probe	✓	88.8	70.1

Table 2. Test accuracy comparison between LaBo and Post-hoc Concept Bottleneck Model (PCBM) on CIFAR-10 and CIFAR-100. "w/ end-to-end" denotes whether the model employs an end-to-end residual predictor from image features to targets.

Method	w/ manual concepts	1	5	Full
CompDL [67]	✓	13.6	33.2	52.6
LaBo (Ours)	✗	**35.1**	**55.7**	**71.8**
Linear Probe	-	28.4	55.4	75.5

Table 3. LaBo and CompDL evaluated on CUB for 1/5/full shots.

Compare with Previous CBM

Method	w/ end-to-end	CIFAR-10	CIFAR-100
PCBM [66]	✗	84.5	56.0
LaBo (Ours)	✗	**87.9**	**69.1**
PCBM-h [66]	✓	87.6	69.9
Linear Probe	✓	88.8	70.1

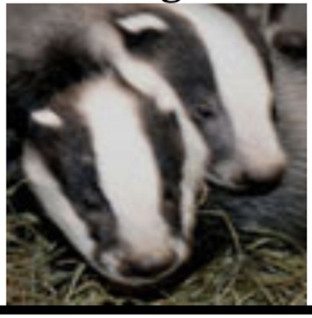
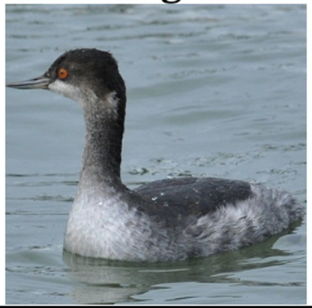
LaBo doesn't rely on black box predictors.
LaBo doesn't require human annotations.

ıoc
00.
ınd
residual predictor from image features to targets.

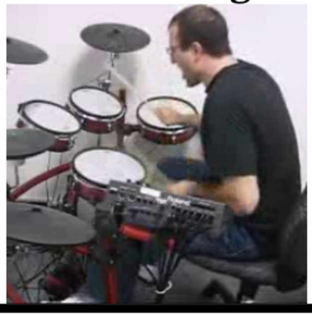
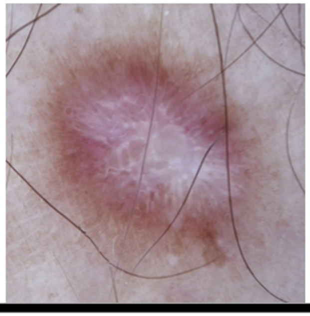
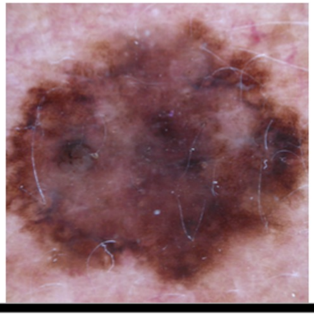
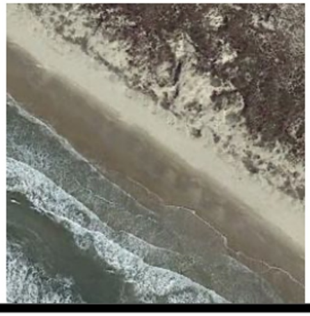
Method	w/ manual concepts	1	5	Full
CompDL [67]	✓	13.6	33.2	52.6
LaBo (Ours)	✗	**35.1**	**55.7**	**71.8**
Linear Probe	-	28.4	55.4	75.5

Table 3. LaBo and CompDL evaluated on CUB for 1/5/full shots.

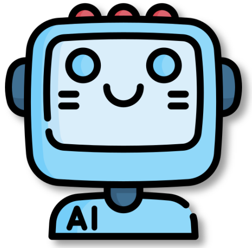
Qualitative Results

	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
ImageNet	**badger** 	1.short legs and long body make it an ex-cellent digger 2. black-and-white striped fur 3. coat is very shaggy	**ant** 	1. black and red stinger 2. small, black insect with six legs 3. long, slender anten-nae that it uses to smell and touch
Food101	**ramen** 	1. garnished with green onions, nori, and other toppings 2. most grocery stores 3. various toppings	**hummus** 	1. chickpeas, tahini, olive oil, garlic, lemon juice 2. made from cooked, mashed chickpeas 3. roasted red peppers
CUB	**eared grebe** 	1. black and white plumage that is strik-ing in the sunlight 2. black body with a long, slender neck 3. red and black bill	**horned lark** 	1. black line running through yellow face 2. head is black with a white horn on each side 3. black horn on each side of their head

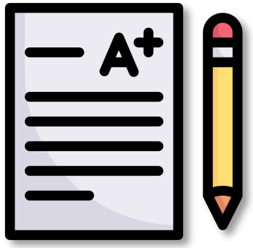
Qualitative Results

UCF-101	**archery** 	1. grip bow tightly in their left hand 2. focused and concentrated on their task 3. keep bow and arrows in safe and dry place when not in use	**drumming** 	1. blur as they fly over the drums 2. sitting on a stool in front of a drum set 3. position the drumstick so it is resting on your index finger
HAM100000	**dermatofibroma** 	1. generally not painful 2. red, brown, or purple in color 3. thin white halo around them	**melanoma** 	1. dark brown or black in color 2. large and dark 3. flesh-colored, brown, or black
RESISC45	**beach** 	1. waves crashing onto the shore 2. few rocks poking out 3. waves are gentle	**railway** 	1. connected by steel rails 2. tramline that is 3 feet wide and runs along the length of the court 3. faint, twinkling line

Conclusion



Leverage the **knowledge of LLM** to build interpretable models (CBMs).

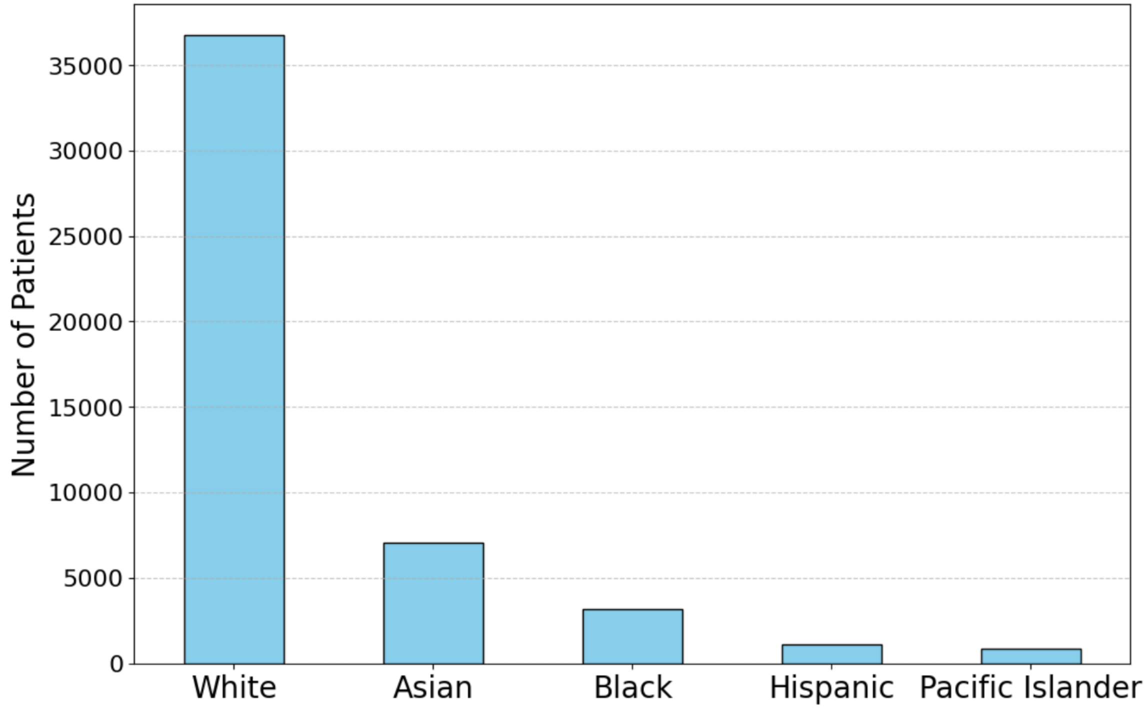


With **vision-language models** (VLMs) and concept selection, interpretable models can achieve **competitive performance** as Black-box.

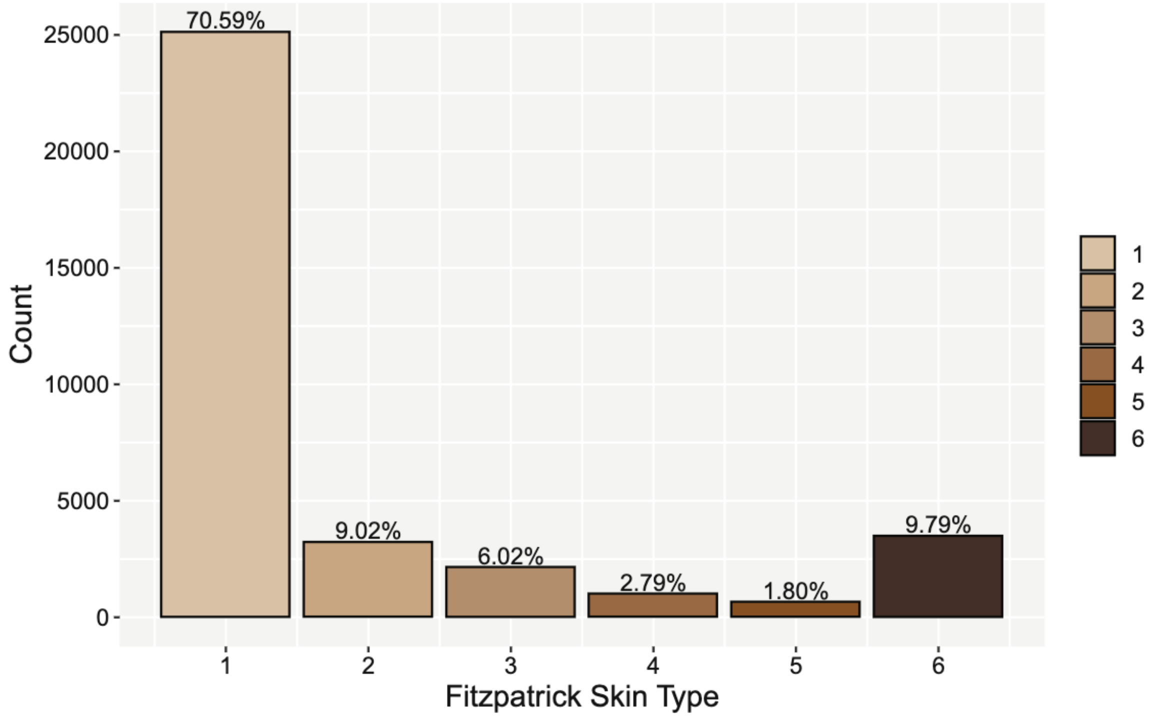
What makes the critical domain more challenging?

The distribution of demographic variables in medical data can be skewed.

Distribution of race in CheXpert [1].



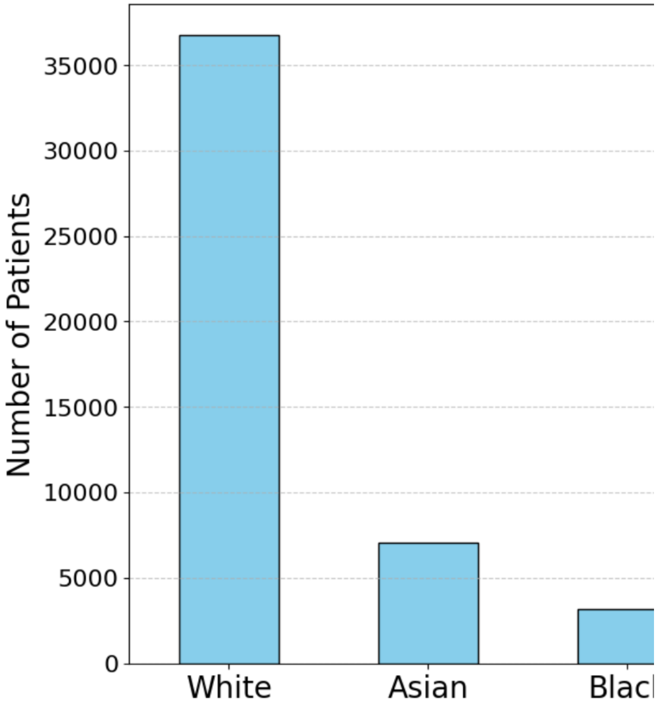
Distribution of skin colors in ISIC [2, 3].



[1] Irvin & Rajpurkar et al. AAAI. 2019; [2] Codella et al. ISBI. 2018; [3] Bevan et al., DART. 2022.

The distribution of … can be skewed.

Distribution of race i… colors in ISIC [2, 3].

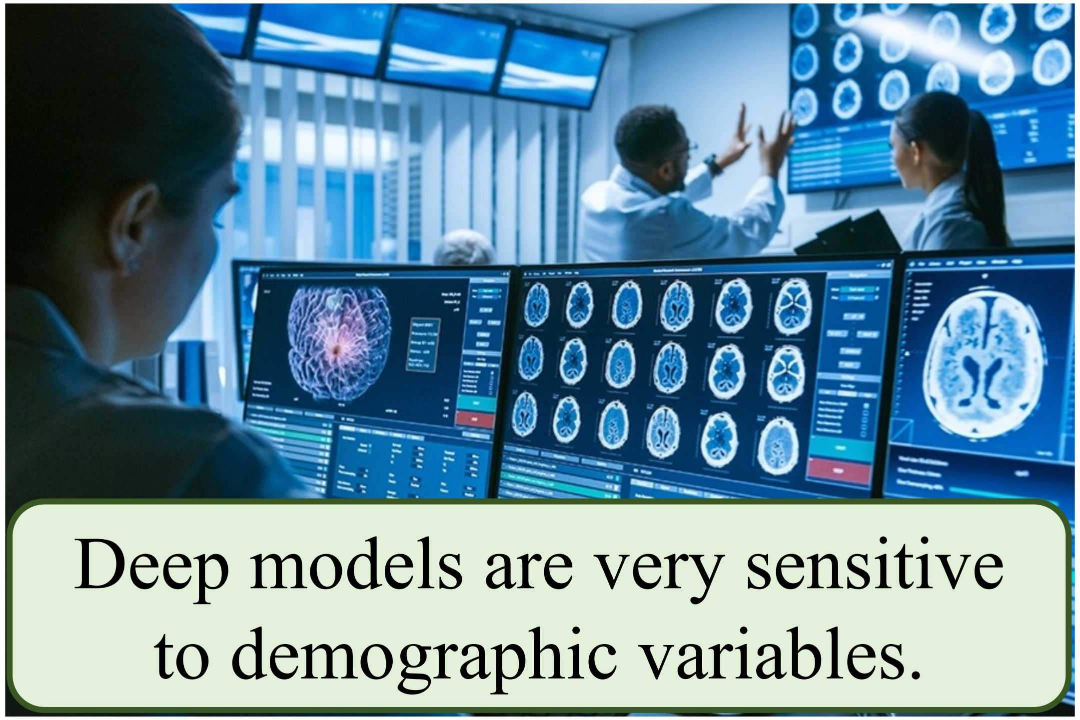


MIT News
ON CAMPUS AND AROUND THE WORLD
SUBSCRIBE

Artificial intelligence predicts patients' race from their medical images

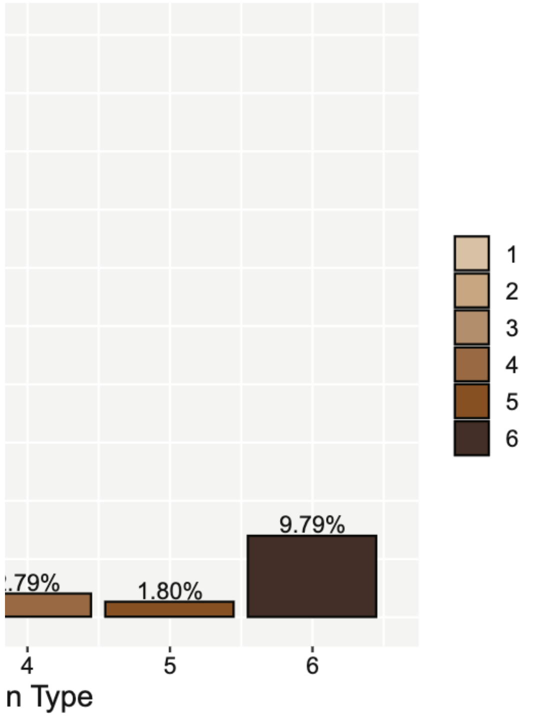
Study shows AI can identify self-reported race from medical images that contain no indications of race detectable by human experts.

Rachel Gordon | MIT CSAIL
May 20, 2022



Deep models are very sensitive to demographic variables.

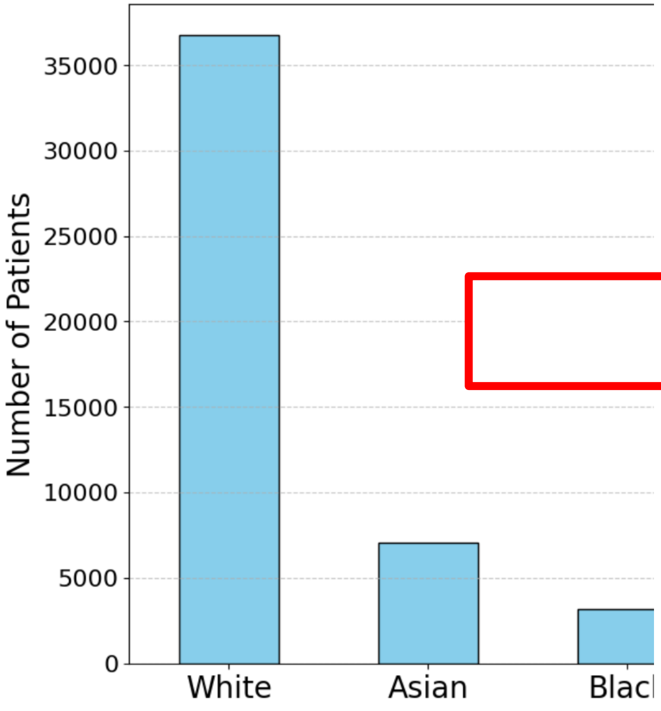
Gichoya et al. Lancet. 2022. Lancet.



[1] Irvin & Rajpurkar et al. AAAI. 2019; [2] Codella et al. ISBI. 2018; [3] Bevan et al., DART. 2022.

The distribution of … can be skewed.

Distribution of race i… colors in ISIC [2, 3].



MIT News

ON CAMPUS AND AROUND THE WORLD

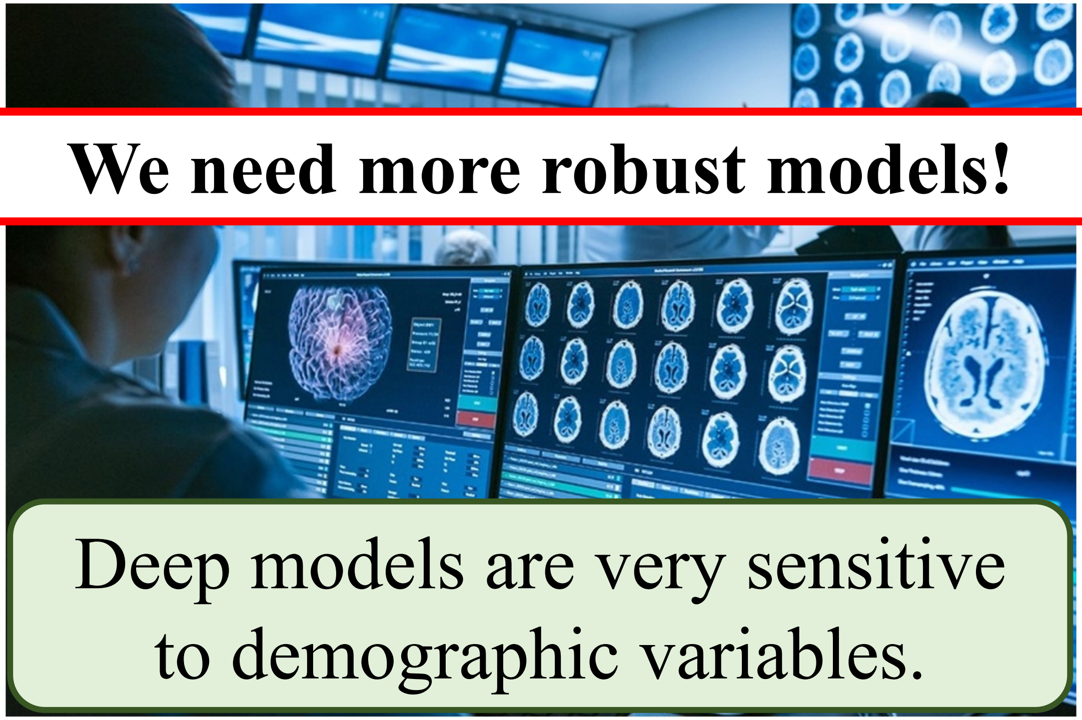
SUBSCRIBE

Artificial intelligence predicts patients' race from their medical images

Study shows AI can identify self-reported race from medical images that contain no indications of race detectable by human experts.

Rachel Gordon | MIT CSAIL

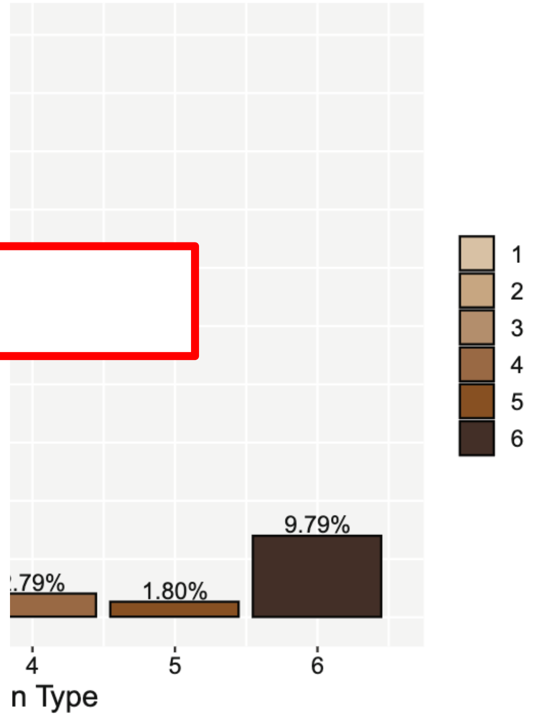
May 20, 2022



We need more robust models!

Deep models are very sensitive to demographic variables.

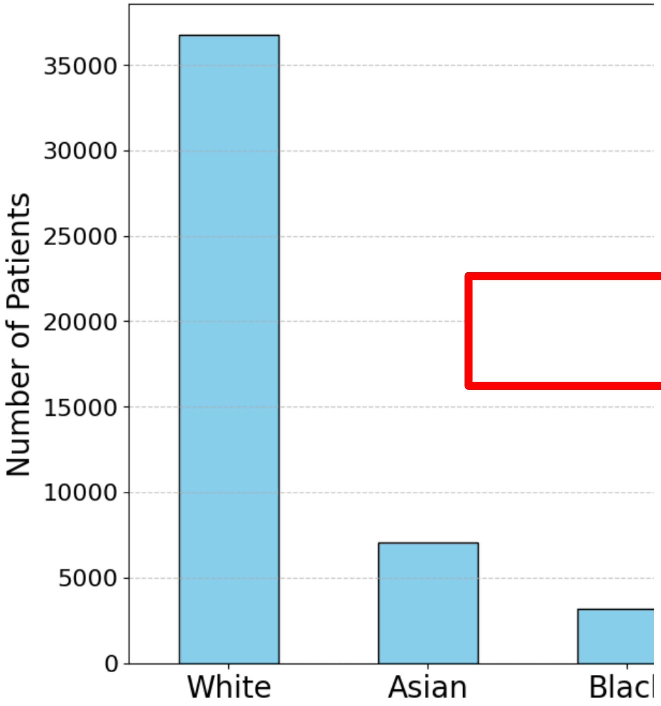
Gichoya et al. Lancet. 2022. Lancet.



[1] Irvin & Rajpurkar et al. AAAI. 2019; [2] Codella et al. ISBI. 2018; [3] Bevan et al., DART. 2022.

The distribution of ... can be skewed.

Distribution of race i... colors in ISIC [2, 3].

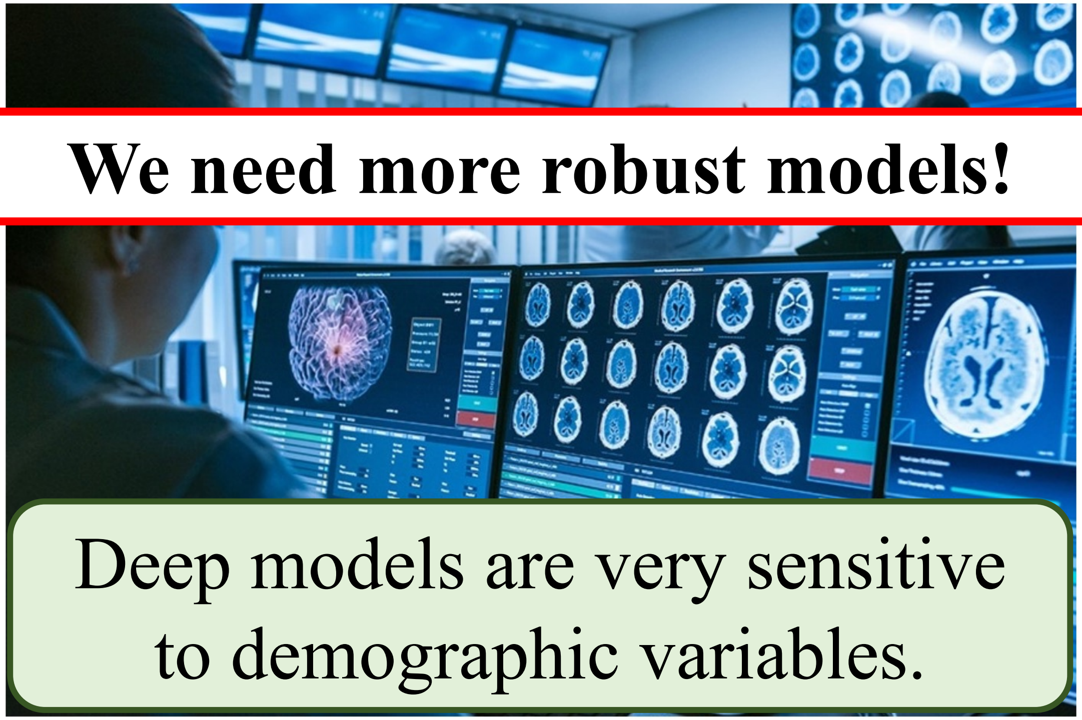


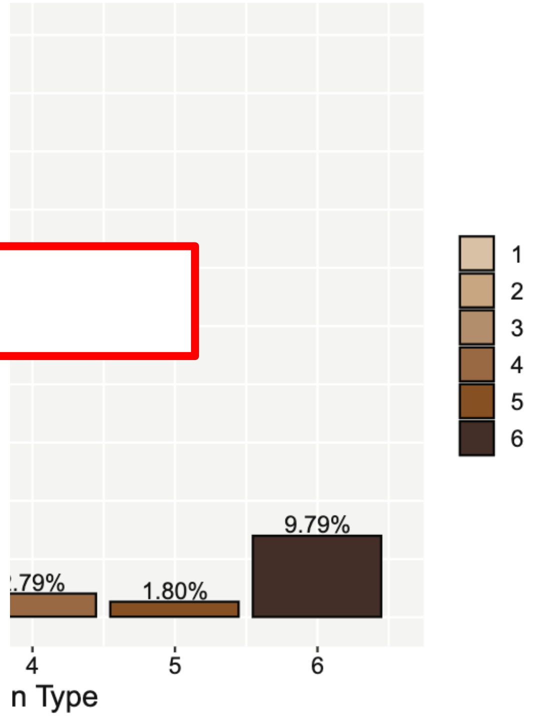
MIT News
ON CAMPUS AND AROUND THE WORLD
SUBSCRIBE

Artificial intelligence predicts patients' race from their medical images

Study shows AI can identify self-reported race from medical images that contain no indications of race detectable by human experts.

Rachel Gordon | MIT CSAIL
May 20, 2022





We need more robust models!

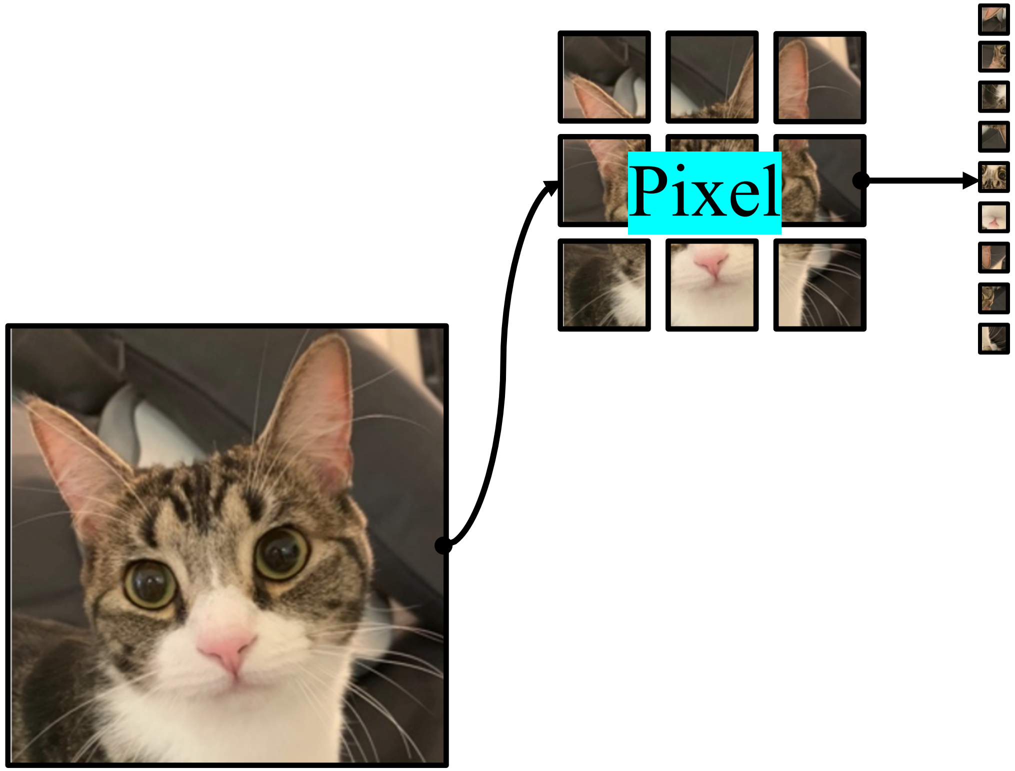
Deep models are very sensitive to demographic variables.

Gichoya et al. Lancet. 2022. Lancet.

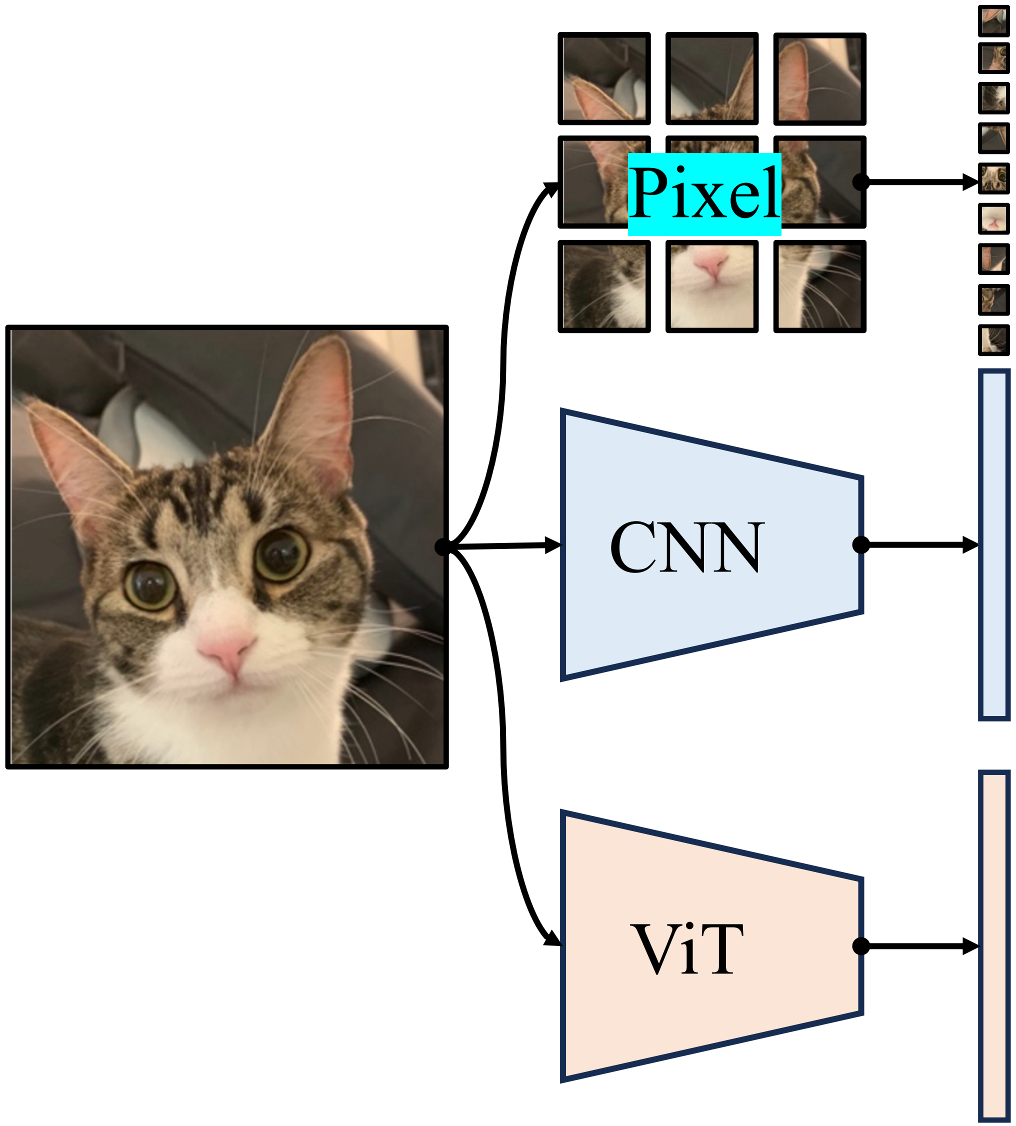
[1] Irvin & Rajpurkar et al. AAAI. 2019; [2] Codella et al. ISBI. 2018; [3] Bevan et al., DART. 2022.

Deep models have good priors for the **general domain**.

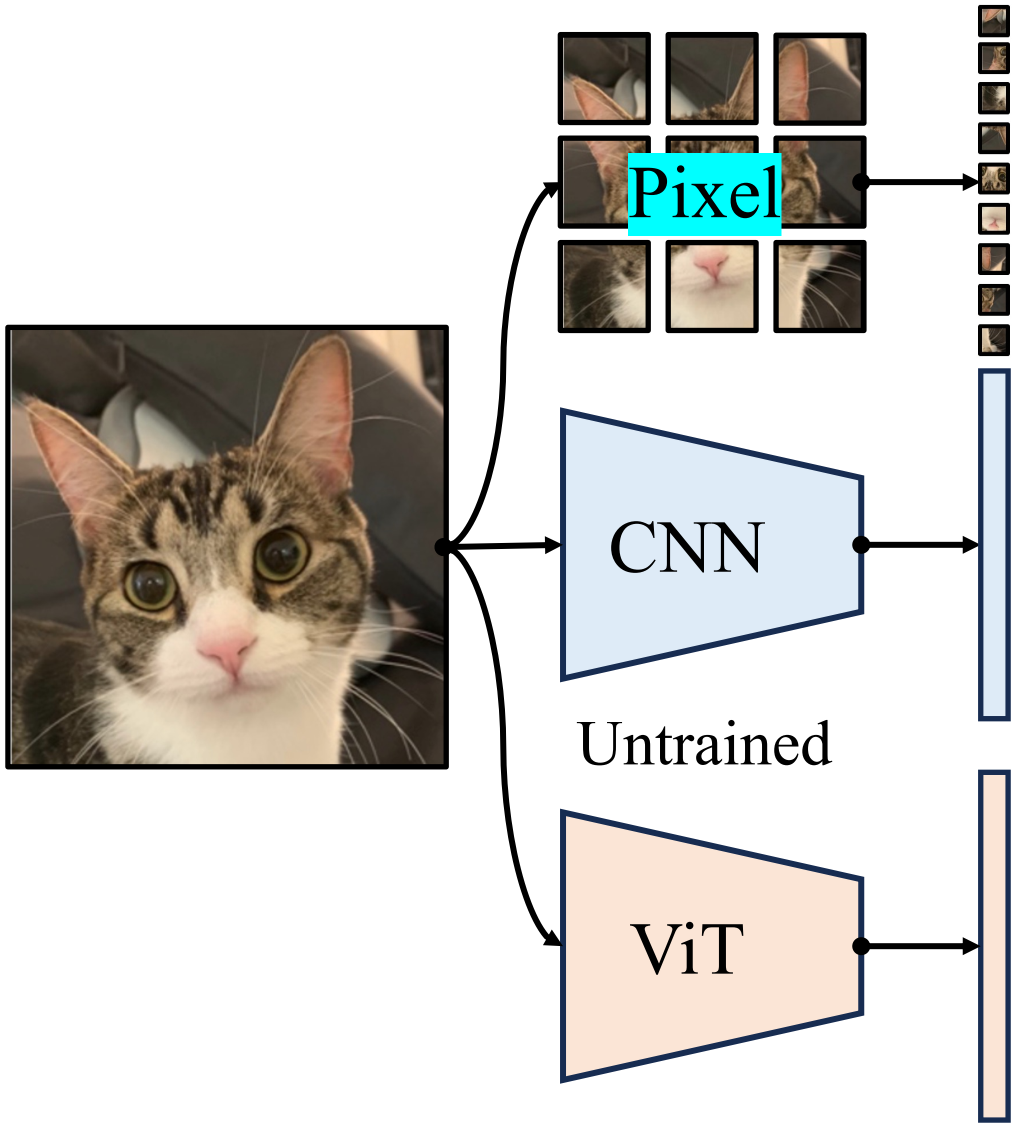
Deep models have good priors for the **general domain**.



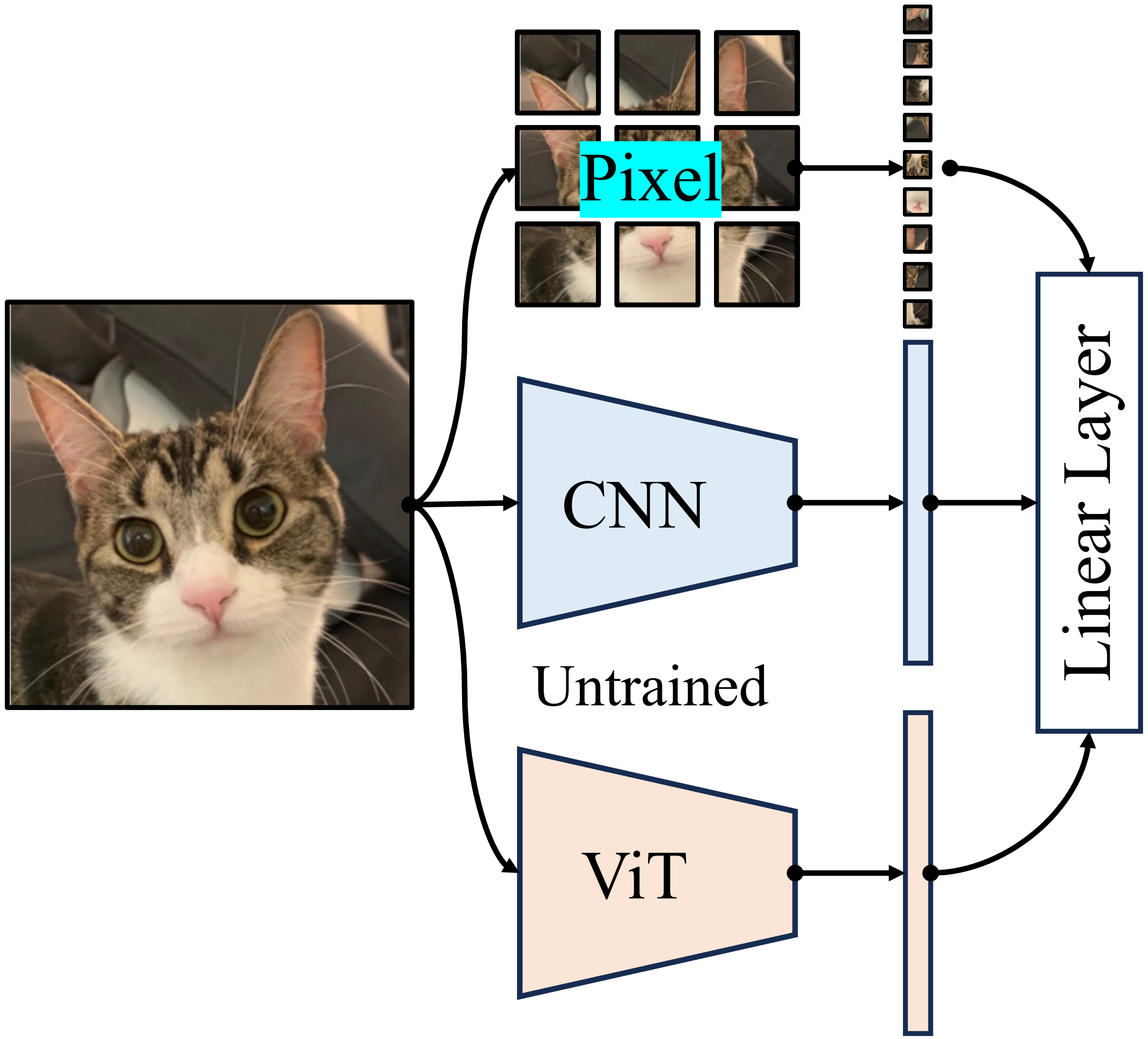
Deep models have good priors for the **general domain**.



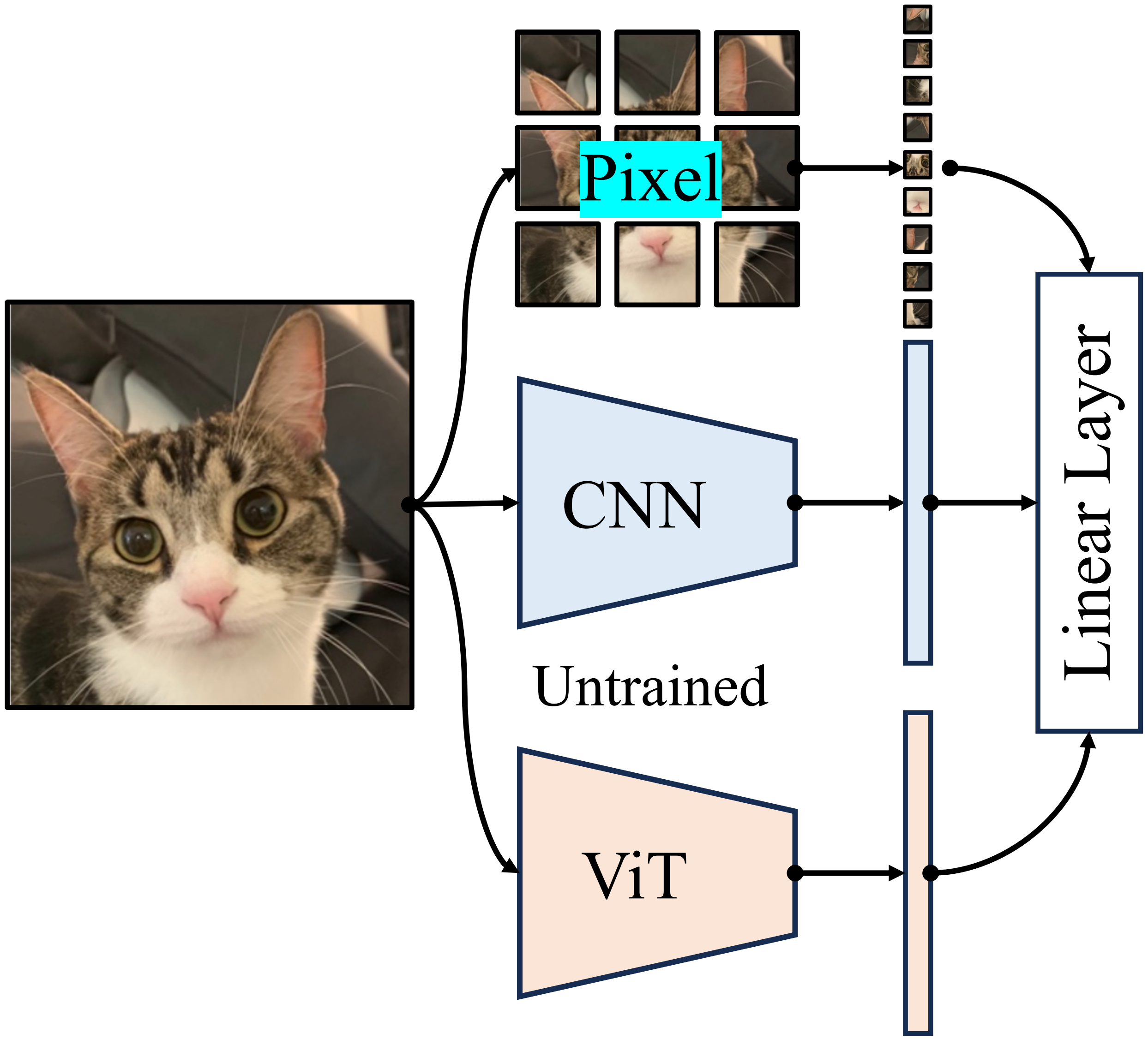
Deep models have good priors for the **general domain**.

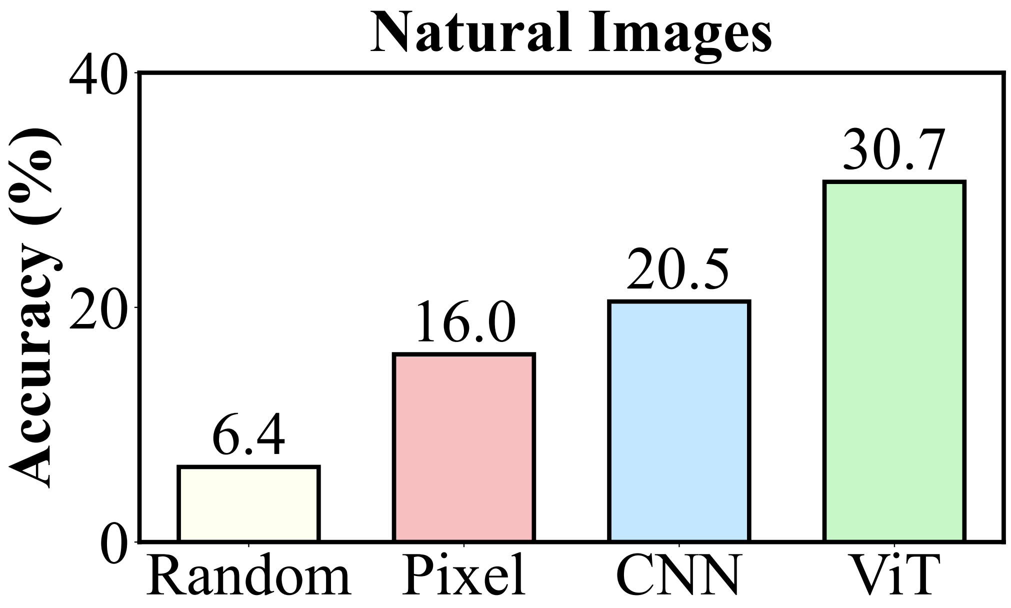


Deep models have good priors for the **general domain**.



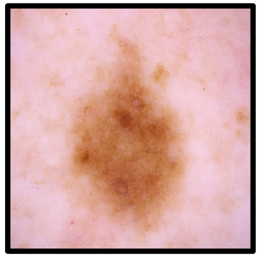
Deep models have good priors for the **general domain**.

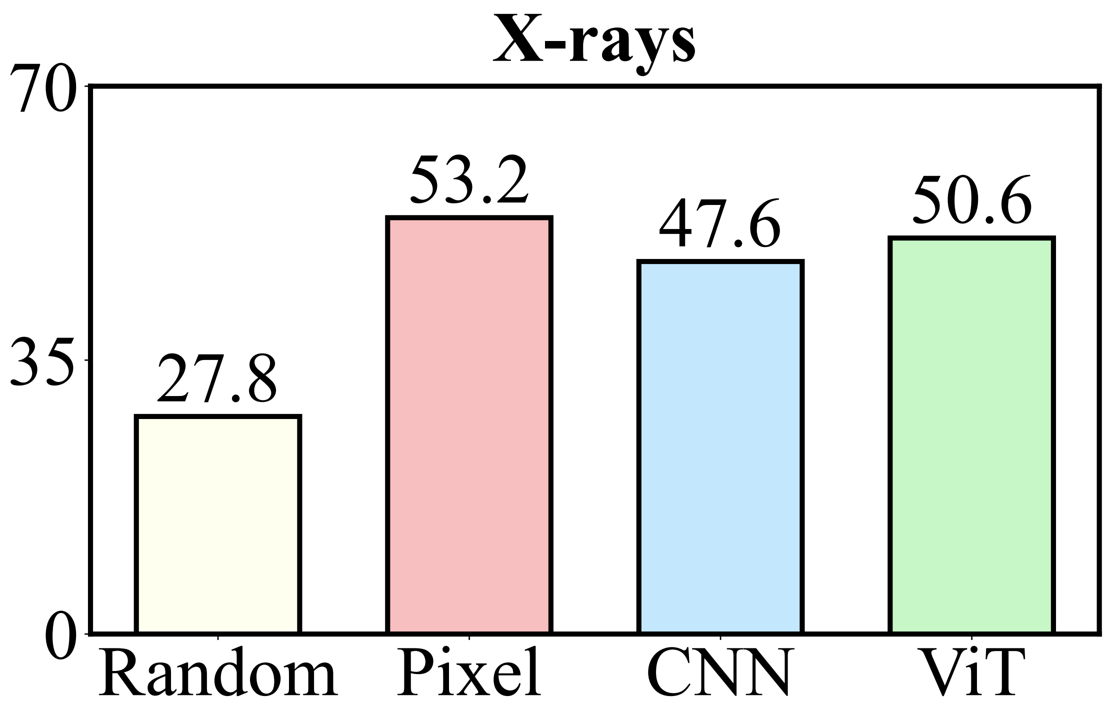




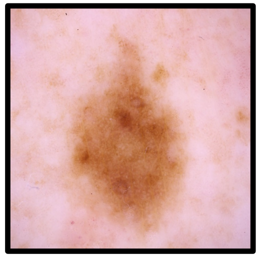
Datasets used: CIFAR-10, STL-10, ImageNet, Flower-101, Food-101

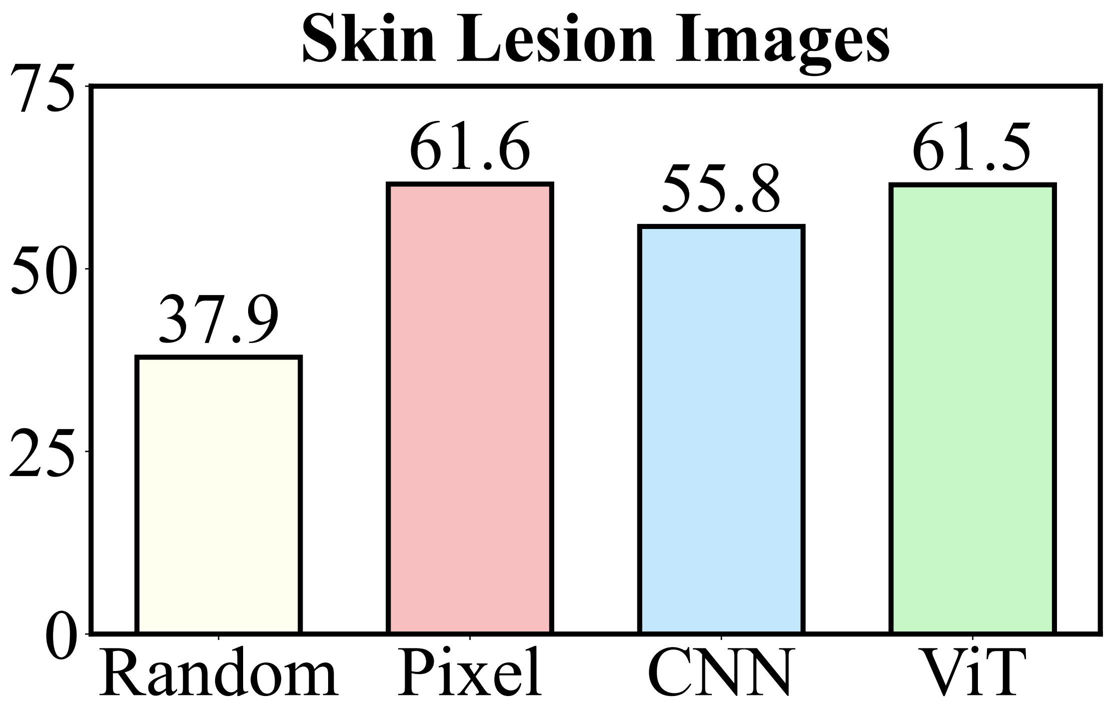

UPRIGHT





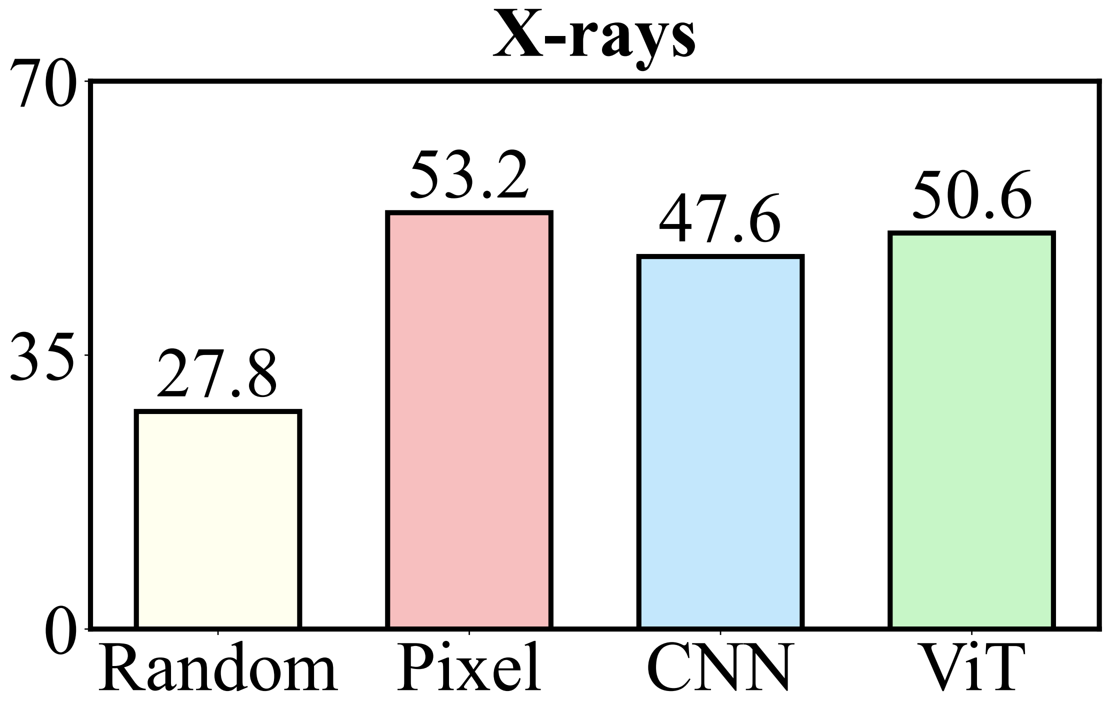
X-ray Datasets: Pneumonia, COVID-QU, NIH-CXR, Open-I, VinDr-CXR.



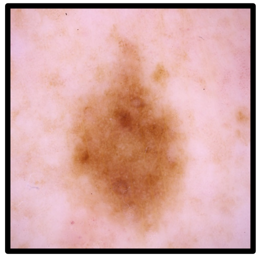


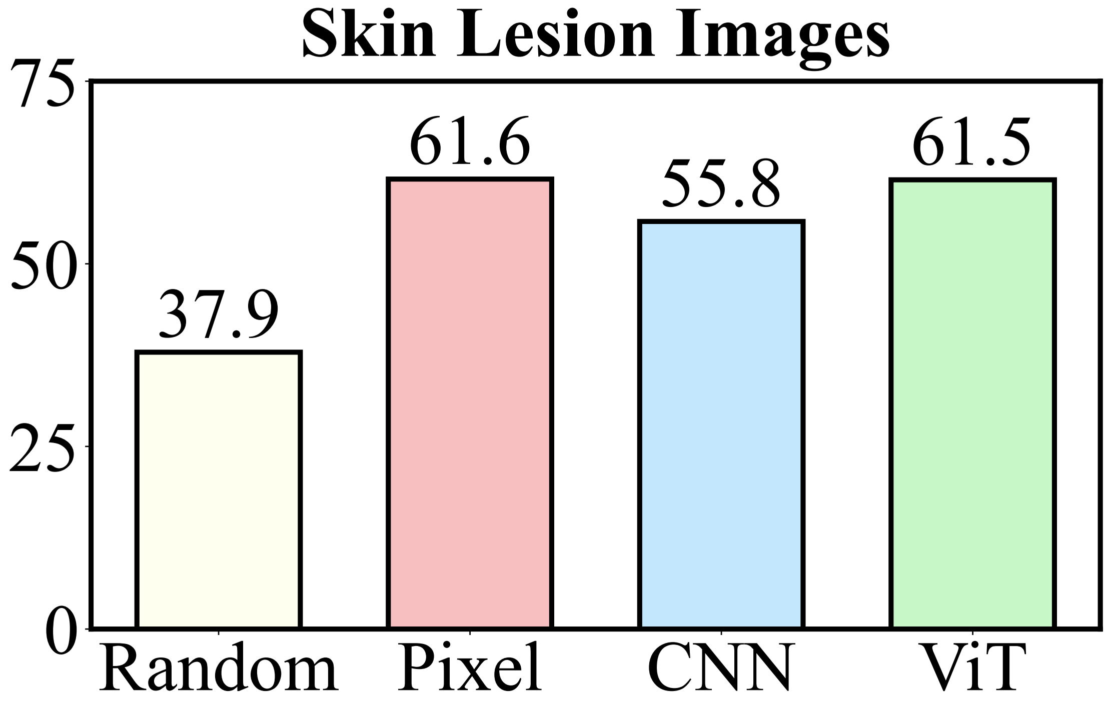
Skin Lesion Datasets: HAM10000, BCN20000, PAD-UFS-20, Melanoma, UWaterloo.

Deep models **don’t have** good priors for the **medical domain**.



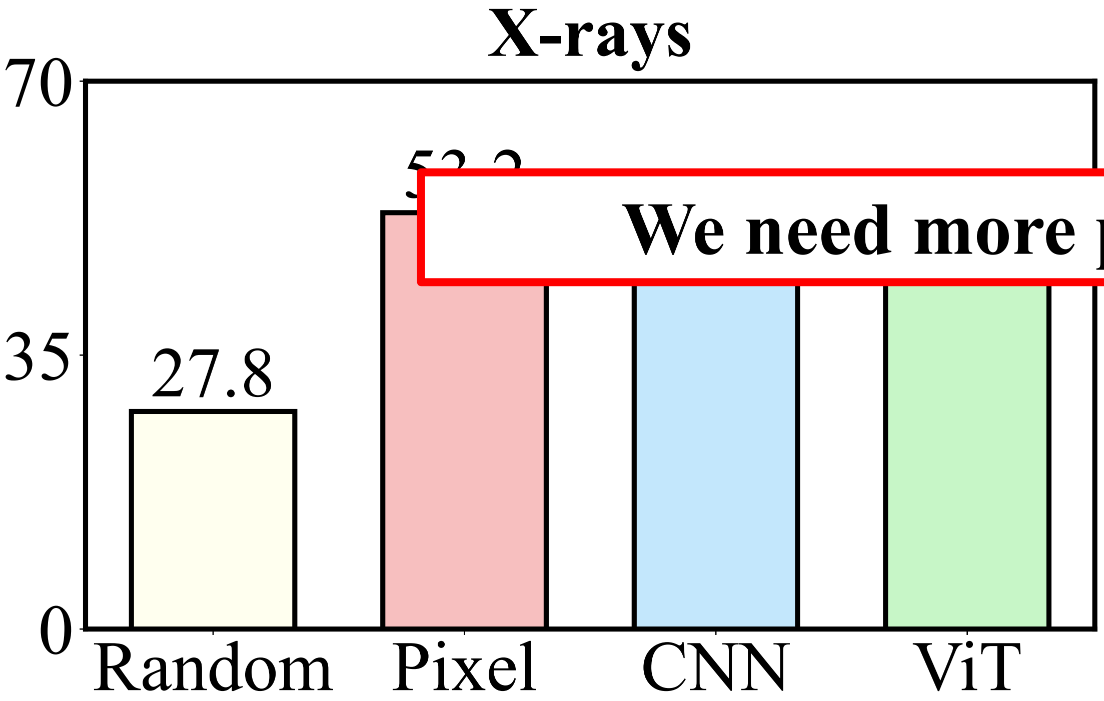
X-ray Datasets: Pneumonia, COVID-QU, NIH-CXR, Open-I, VinDr-CXR.

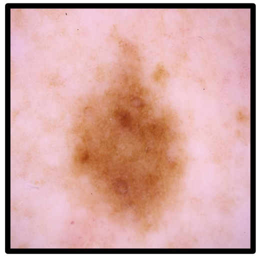


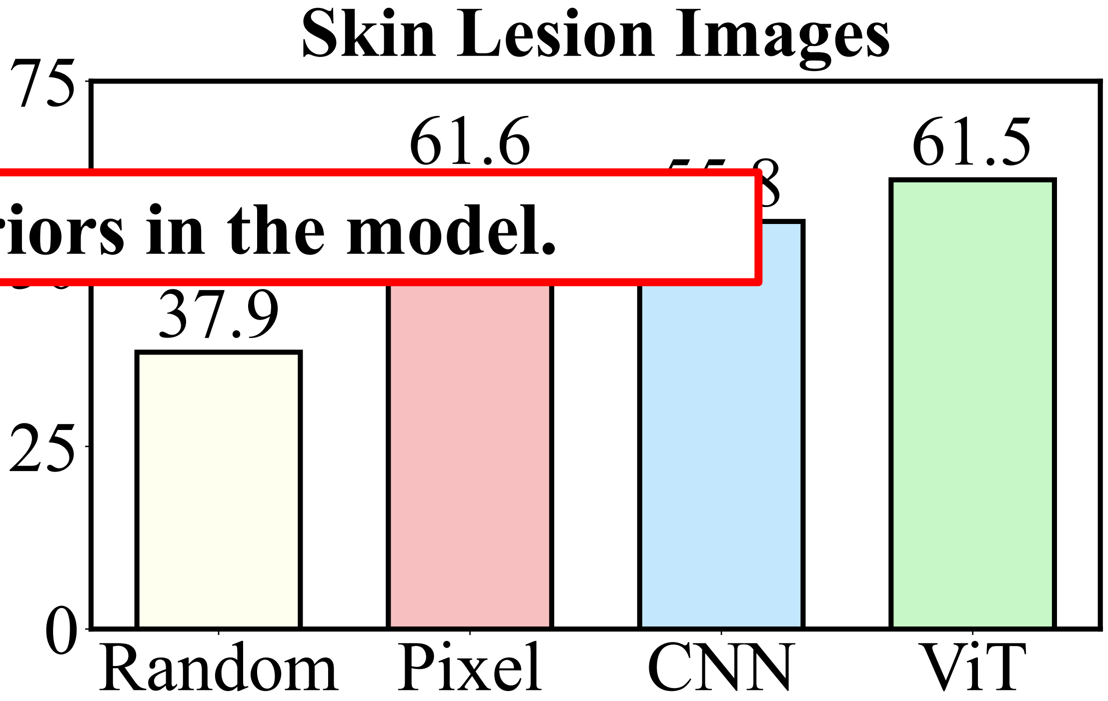


Skin Lesion Datasets: HAM10000, BCN20000, PAD-UFS-20, Melanoma, UWaterloo.

Deep models **don’t have** good priors for the **medical domain**.







We need more priors in the model.

X-ray Datasets: Pneumonia, COVID-QU, NIH-CXR, Open-I, VinDr-CXR.

Skin Lesion Datasets: HAM10000, BCN20000, PAD-UFS-20, Melanoma, UWaterloo.

KnoBo: Knowledge-enhanced Concept Bottlenecks for Interpretable and Robust Medical Image Classification

Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S. Yao,
James C. Gee, Mark Yatskar

Penn Engineering
University of Pennsylvania

Perelman
School of Medicine
University of Pennsylvania

Where to acquire the prior knowledge?

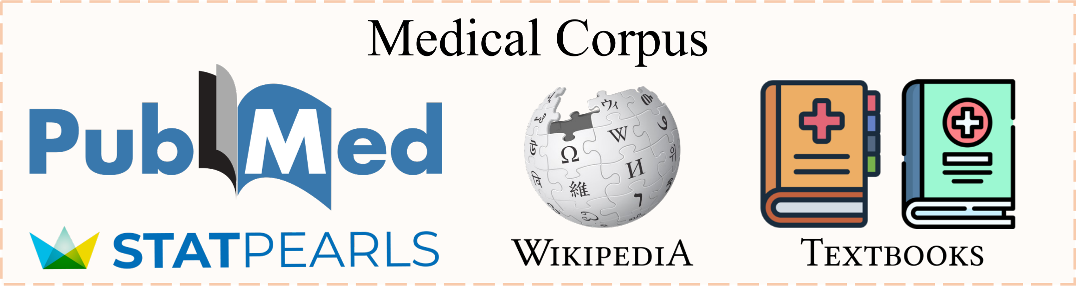
Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?

Query: How to diagnose COVID from X-rays?

Where to acquire the prior knowledge?

Query: How to diagnose COVID from X-rays?



Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?

Query: How to diagnose COVID from X-rays?

Medical Corpus

PubMed

STATPEARLS

WIKIPEDIA

TEXTBOOKS

PubMed 5M Articles, 300M+ paragraphs

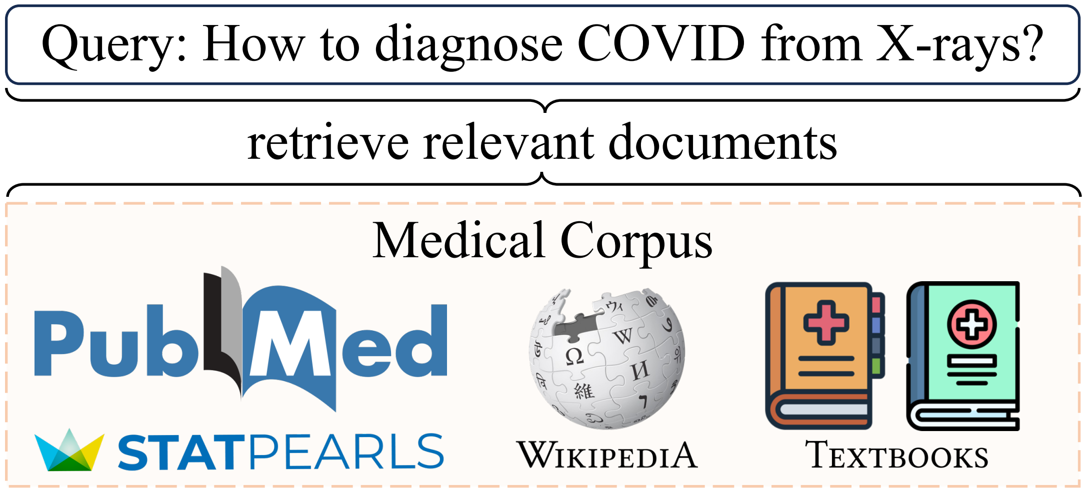
STATPEARLS 9.3K Articles, 301.2K paragraphs

WIKIPEDIA 6.5M Articles, 30.4M paragraphs

TEXTBOOKS 18 Medical Textbooks, 125.8k paragraphs

Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?



PubMed 5M Articles, 300M+ paragraphs

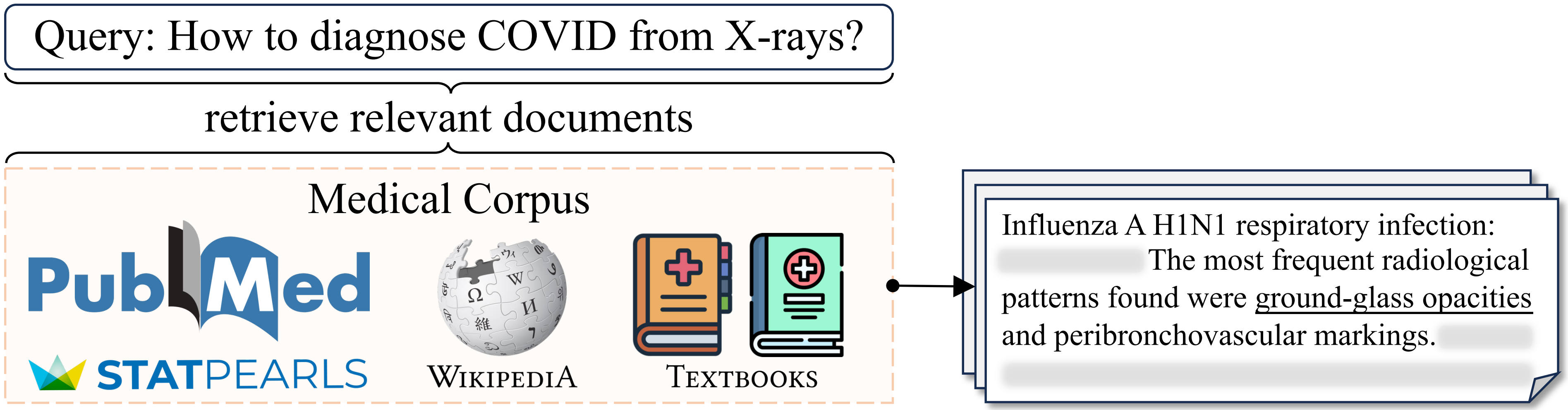
STATPEARLS 9.3K Articles, 301.2K paragraphs

WIKIPEDIA 6.5M Articles, 30.4M paragraphs

TEXTBOOKS 18 Medical Textbooks, 125.8k paragraphs

Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?



PubMed: 5M Articles, 300M+ paragraphs

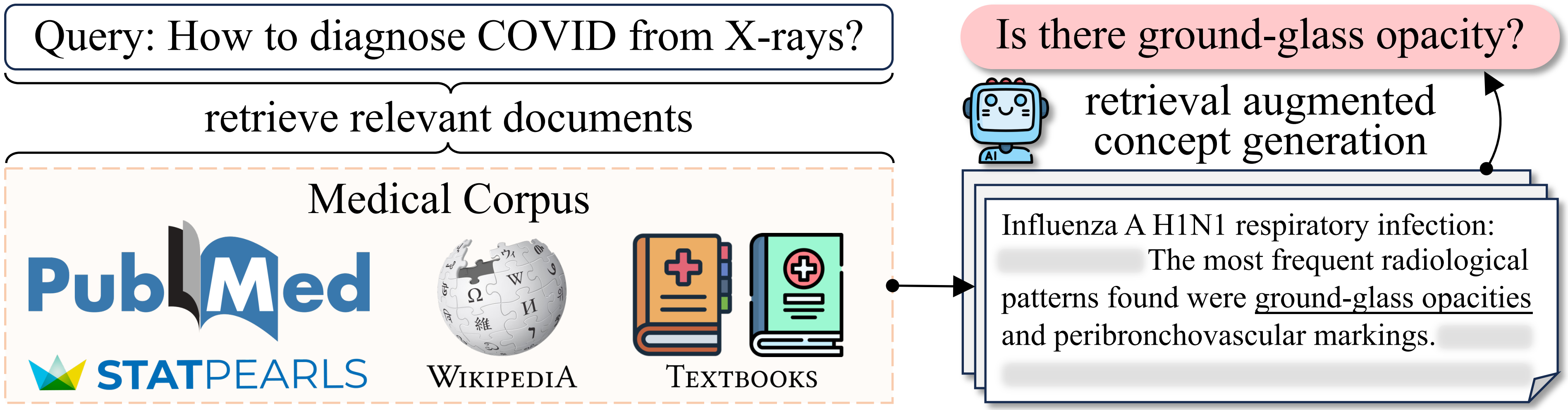
STATPEARLS: 9.3K Articles, 301.2K paragraphs

WIKIPEDIA: 6.5M Articles, 30.4M paragraphs

TEXTBOOKS: 18 Medical Textbooks, 125.8k paragraphs

Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?



PubMed: 5M Articles, 300M+ paragraphs

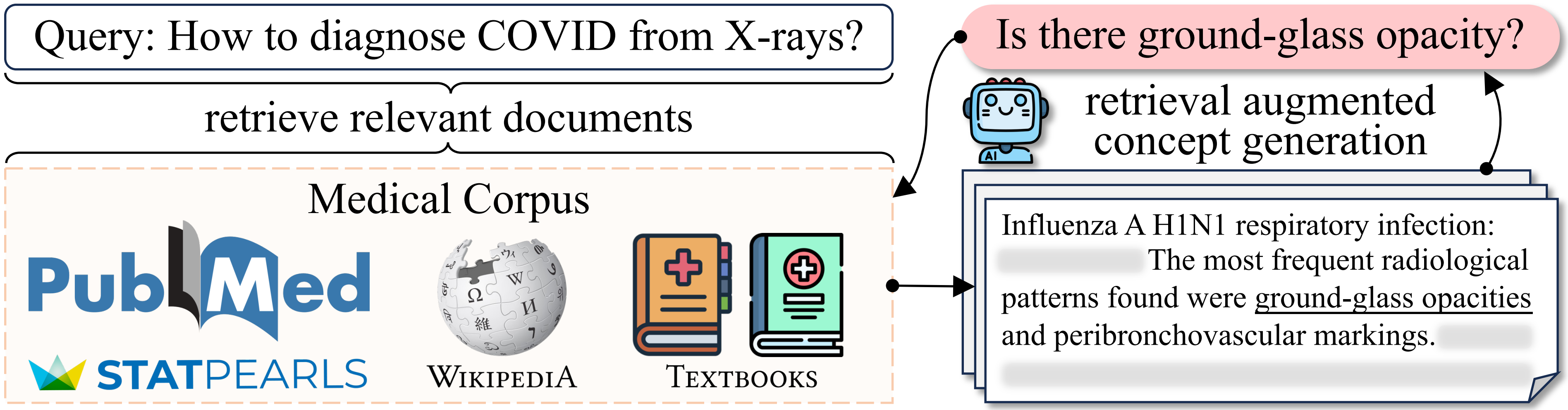
STATPEARLS: 9.3K Articles, 301.2K paragraphs

WIKIPEDIA: 6.5M Articles, 30.4M paragraphs

TEXTBOOKS: 18 Medical Textbooks, 125.8k paragraphs

Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

Where to acquire the prior knowledge?



PubMed: 5M Articles, 300M+ paragraphs

STATPEARLS: 9.3K Articles, 301.2K paragraphs

WIKIPEDIA: 6.5M Articles, 30.4M paragraphs

TEXTBOOKS: 18 Medical Textbooks, 125.8k paragraphs

Xiong et al. Benchmarking Retrieval-Augmented Generation for Medicine. Arxiv. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.

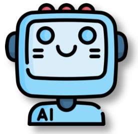
Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.



LLM

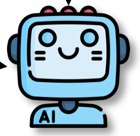
Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.



LLM

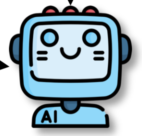
Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.



LLM

Yes

Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

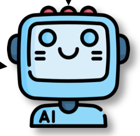
Learn a binary classifier for each concept.

Is there ground-glass opacity?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.



LLM

Yes

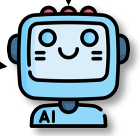
Chung et al. Scaling instruction-finetuned language models. JMLR. 2024.

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.



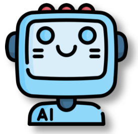
LLM

Yes

Learn a binary classifier for each concept.

Is there ground-glass opacity?

Clinical Report



How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to **ground-glass opacities** on prior CT and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.

LLM

Yes

Learn a binary classifier for each concept.

Is there ground-glass opacity?

Clinical Report

Yes

positive report

No

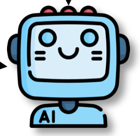
negative report

How to ground the knowledge?

Is there ground-glass opacity?

Paired Clinical Reports

Redemonstration of subtle posterior lung base densities corresponding to prior CT **ground-glass opacities** and likely representing aspiration do not appear worsened. Tiny bilateral pleural effusions.

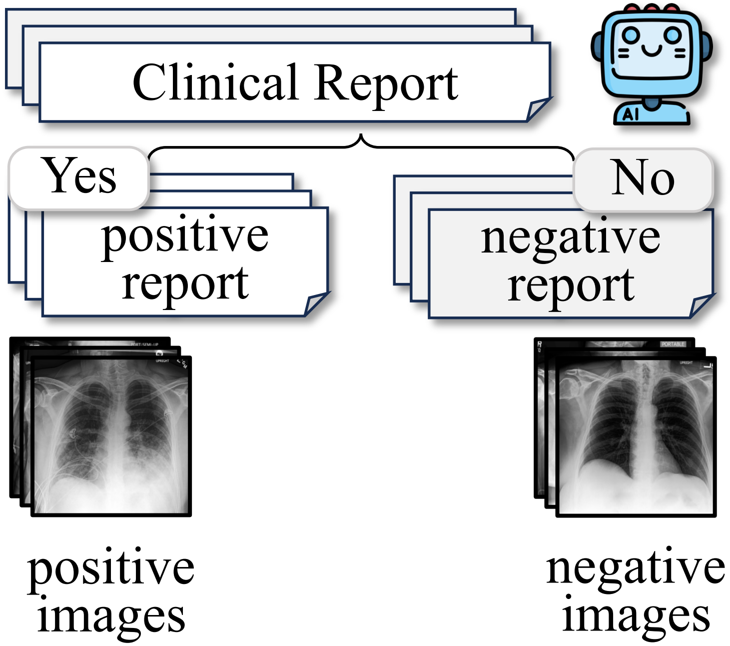


LLM

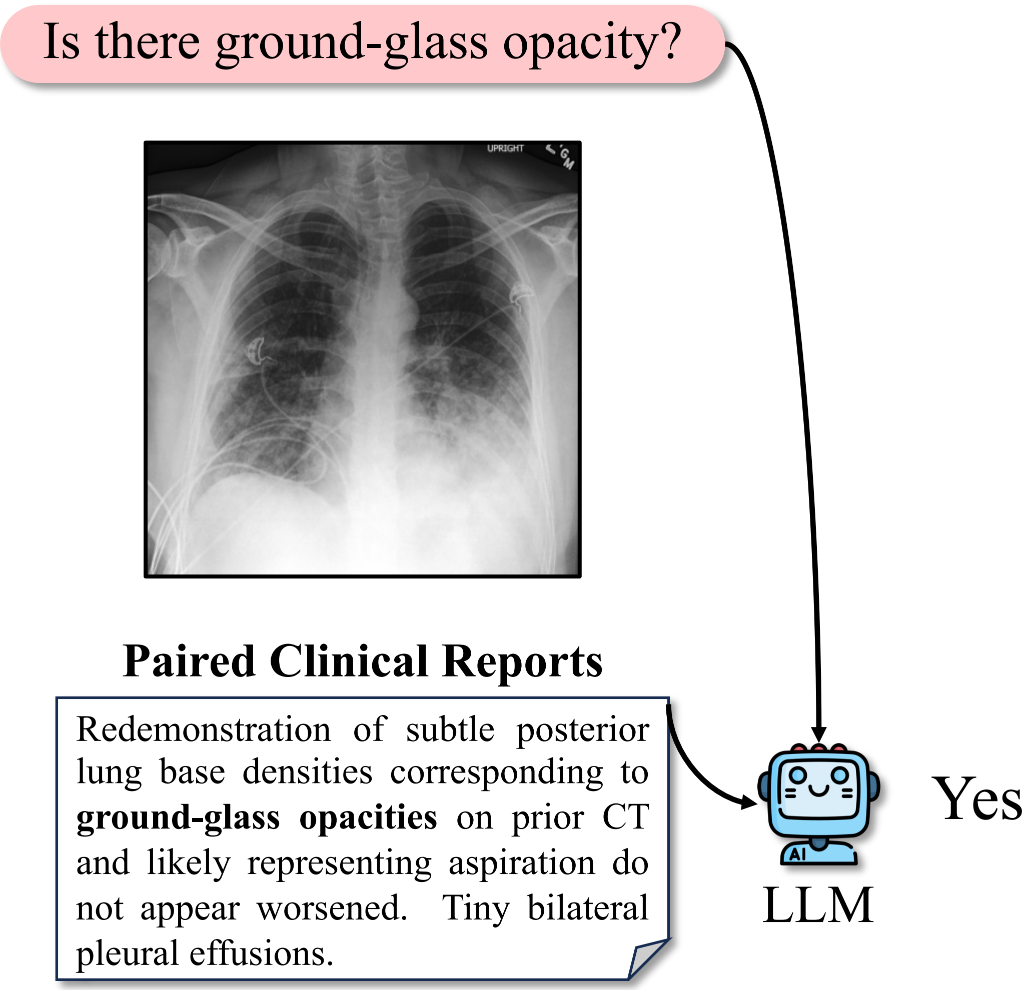
Yes

Learn a binary classifier for each concept.

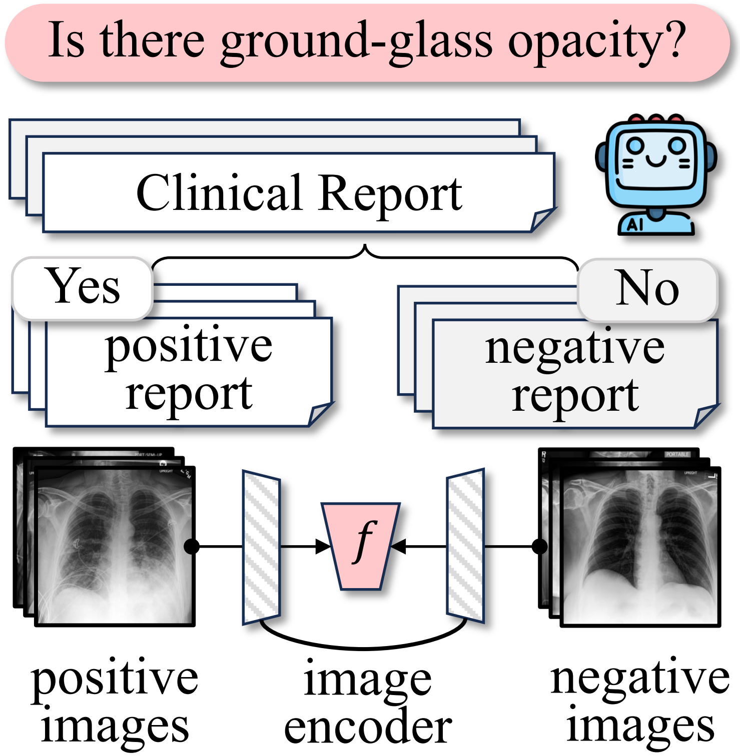
Is there ground-glass opacity?



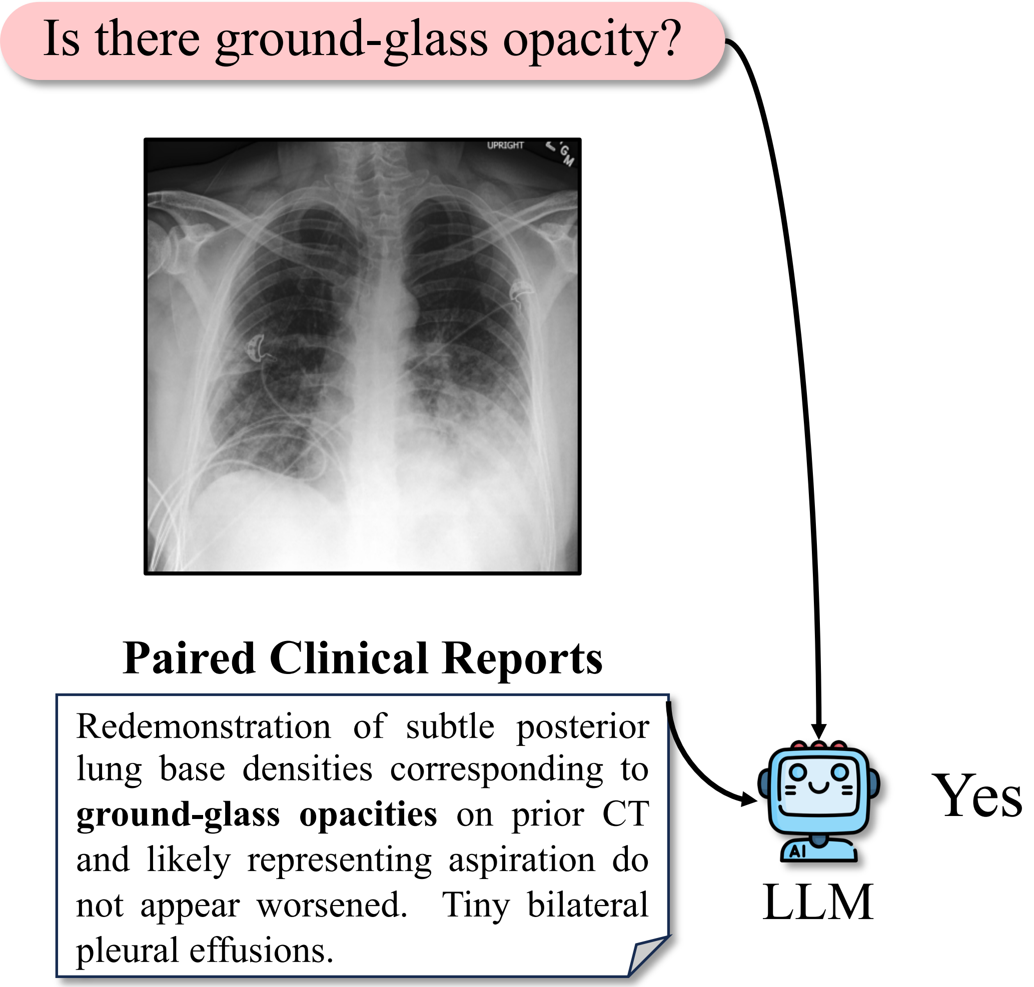
How to ground the knowledge?



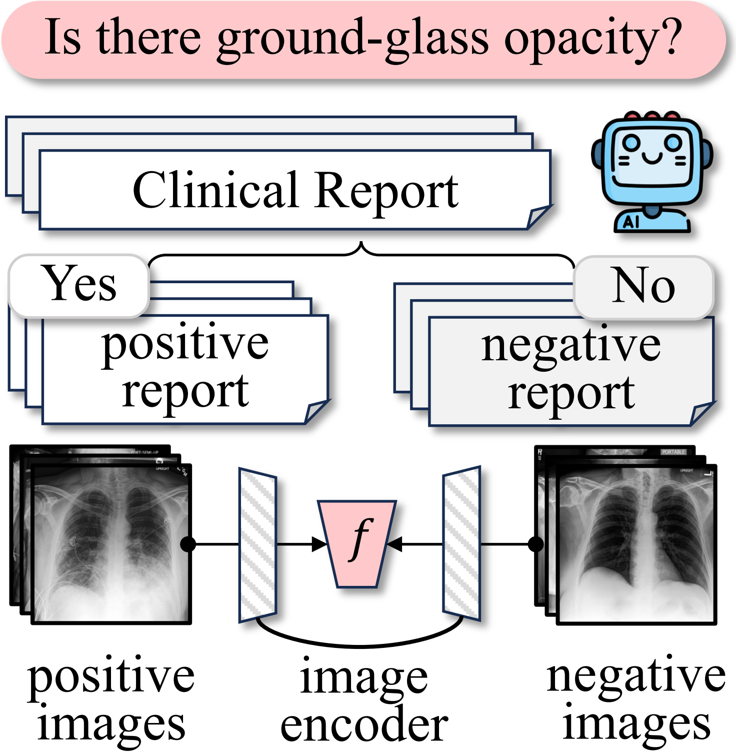
Learn a binary classifier for each concept.



How to ground the knowledge?



Learn a binary classifier for each concept.



Convert text instances into image instances.

Grounding Comparison

Query: Are lung fields clear on both sides?

CLIP **Ours (w/ knowledge grounding)**

Grounding Comparison

Query: Are lung fields clear on both sides?

CLIP

Ours (w/ knowledge grounding)

Top-3 clear

Grounding Comparison

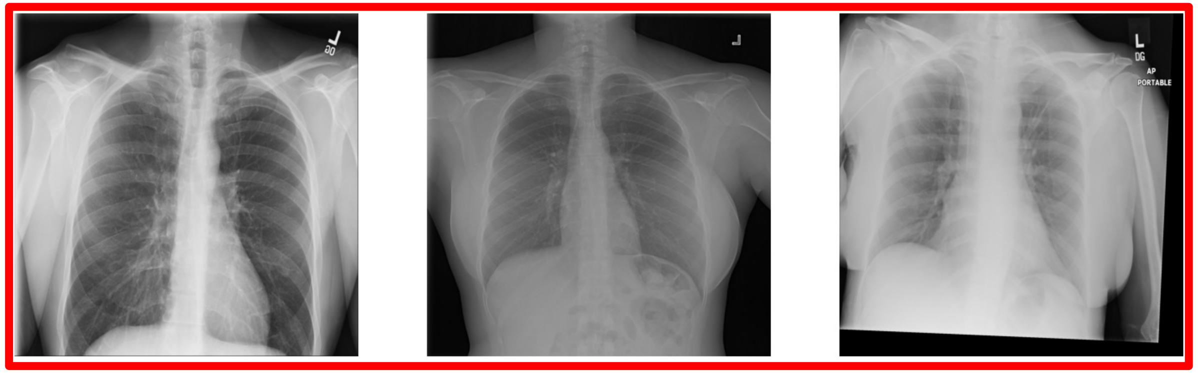
Query: Are lung fields clear on both sides?

CLIP

Top-3 clear

Ours (w/ knowledge grounding)

Top-3 clear



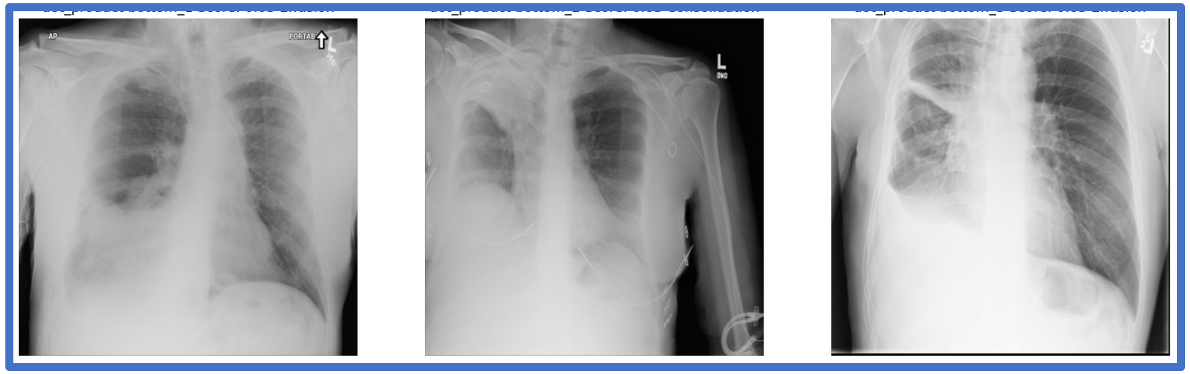
Grounding Comparison

Query: Are lung fields clear on both sides?

CLIP

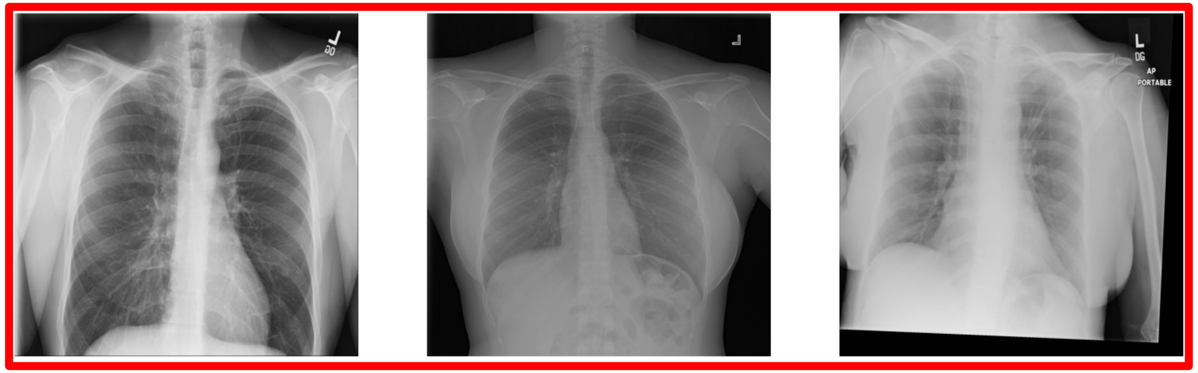
Top-3 clear

Bottom-3 not clear



Ours (w/ knowledge grounding)

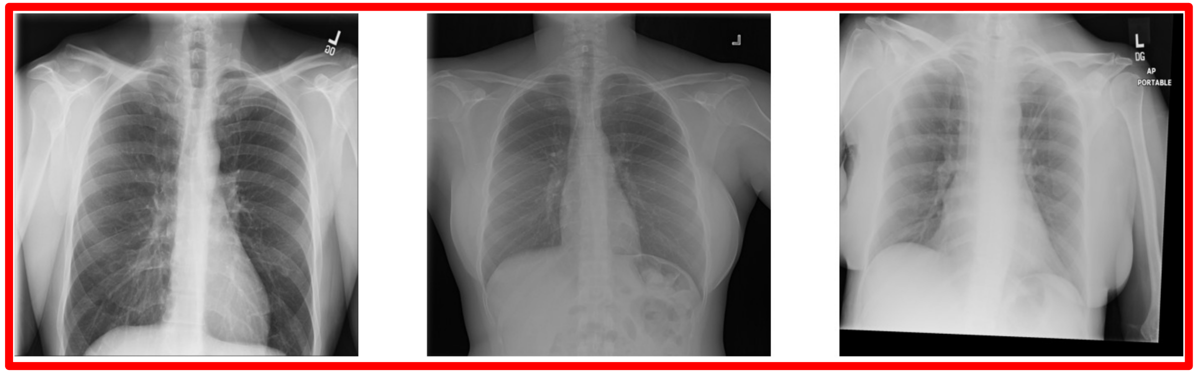
Top-3 clear



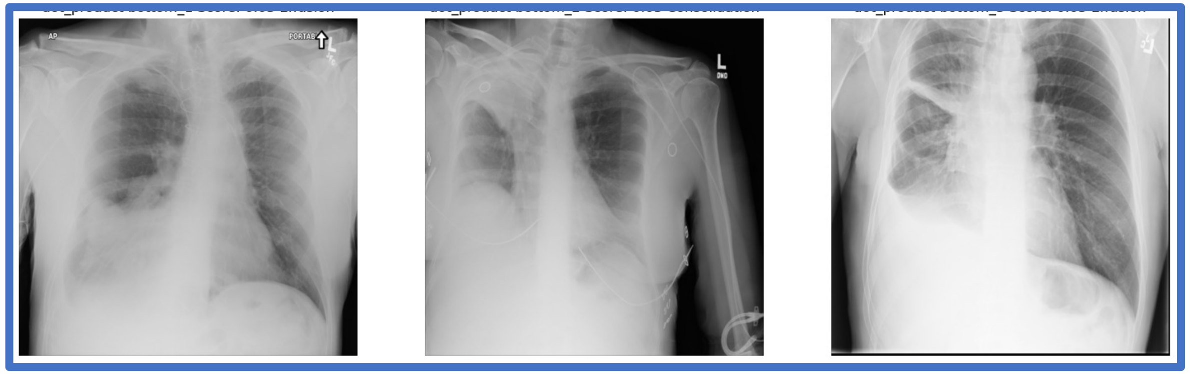
Grounding Comparison

Query: Are lung fields clear on both sides?

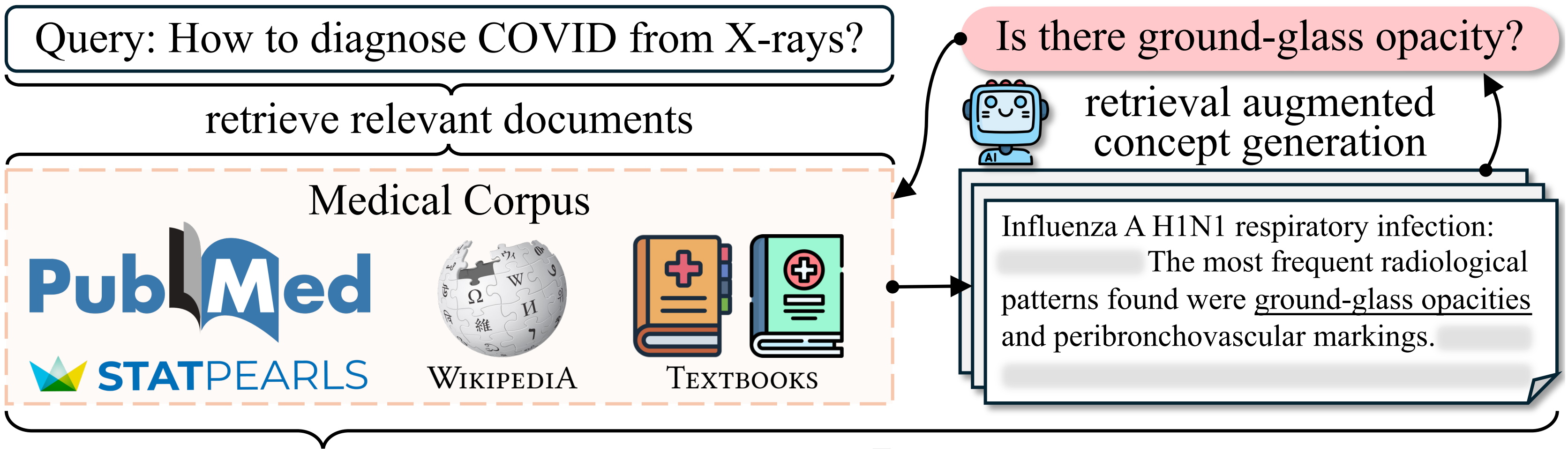
CLIP	**Ours (w/ knowledge grounding)**
Top-3 clear	Top-3 clear



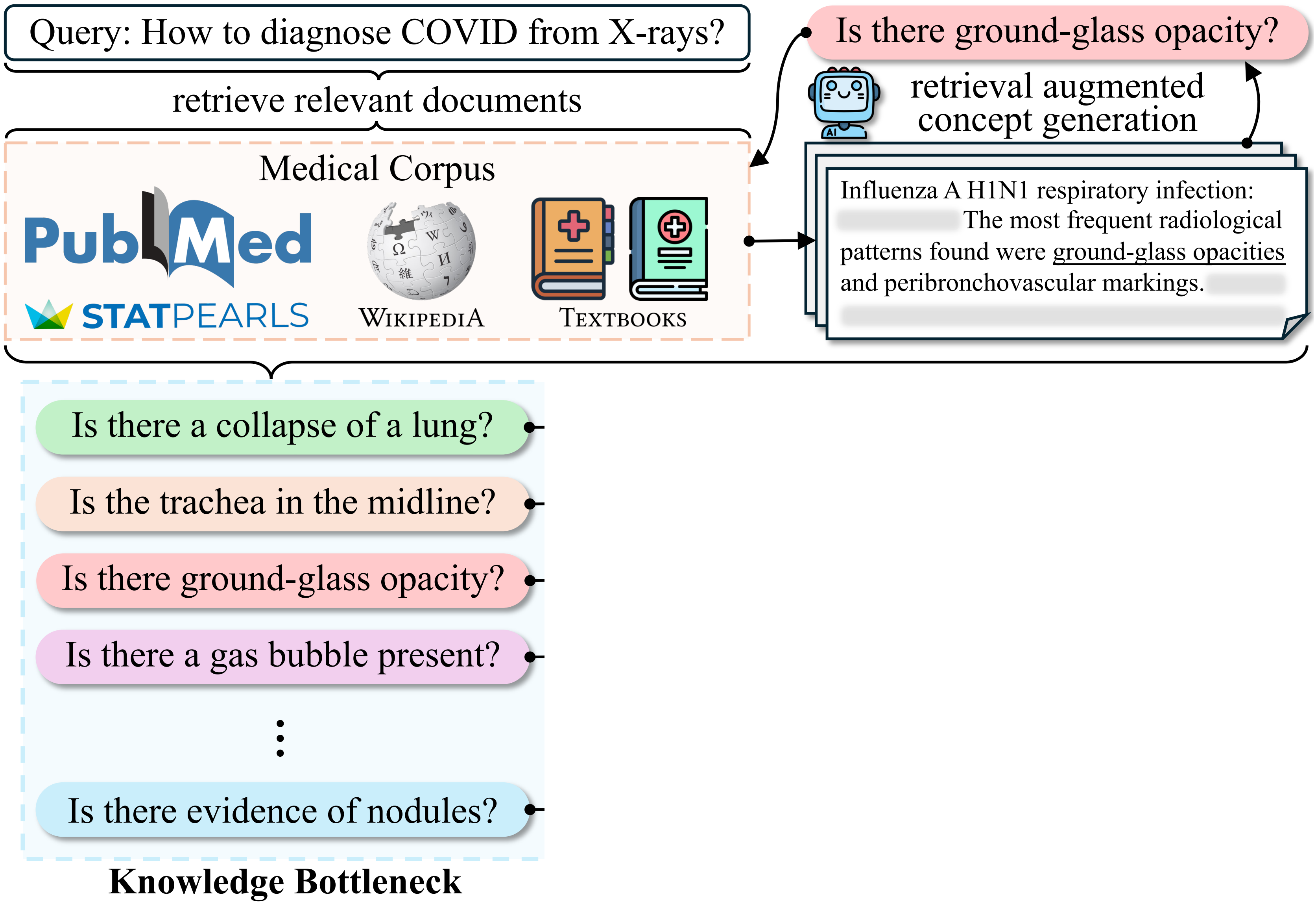
Bottom-3 not clear	Bottom-3 not clear



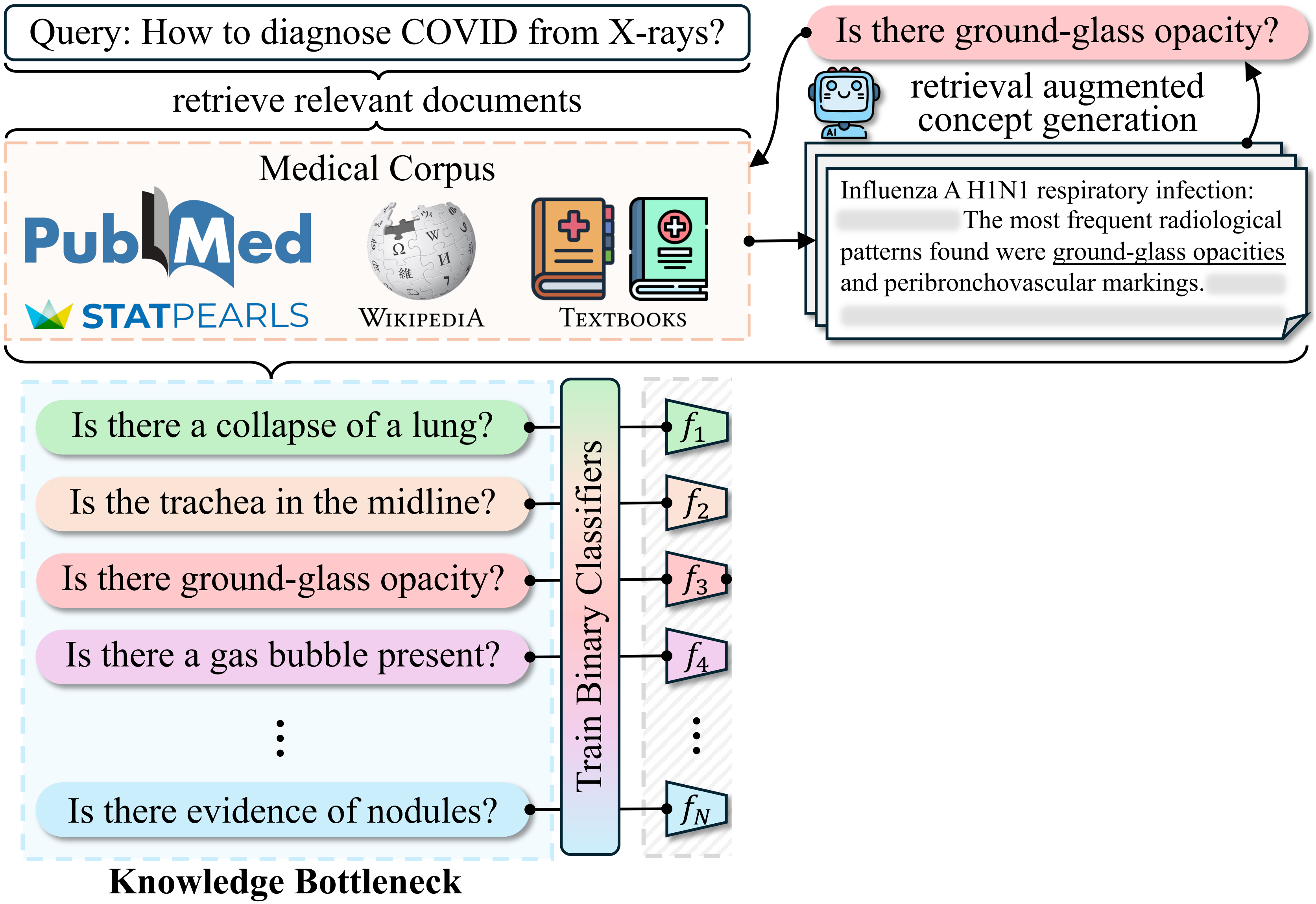
KnoBo System Overview



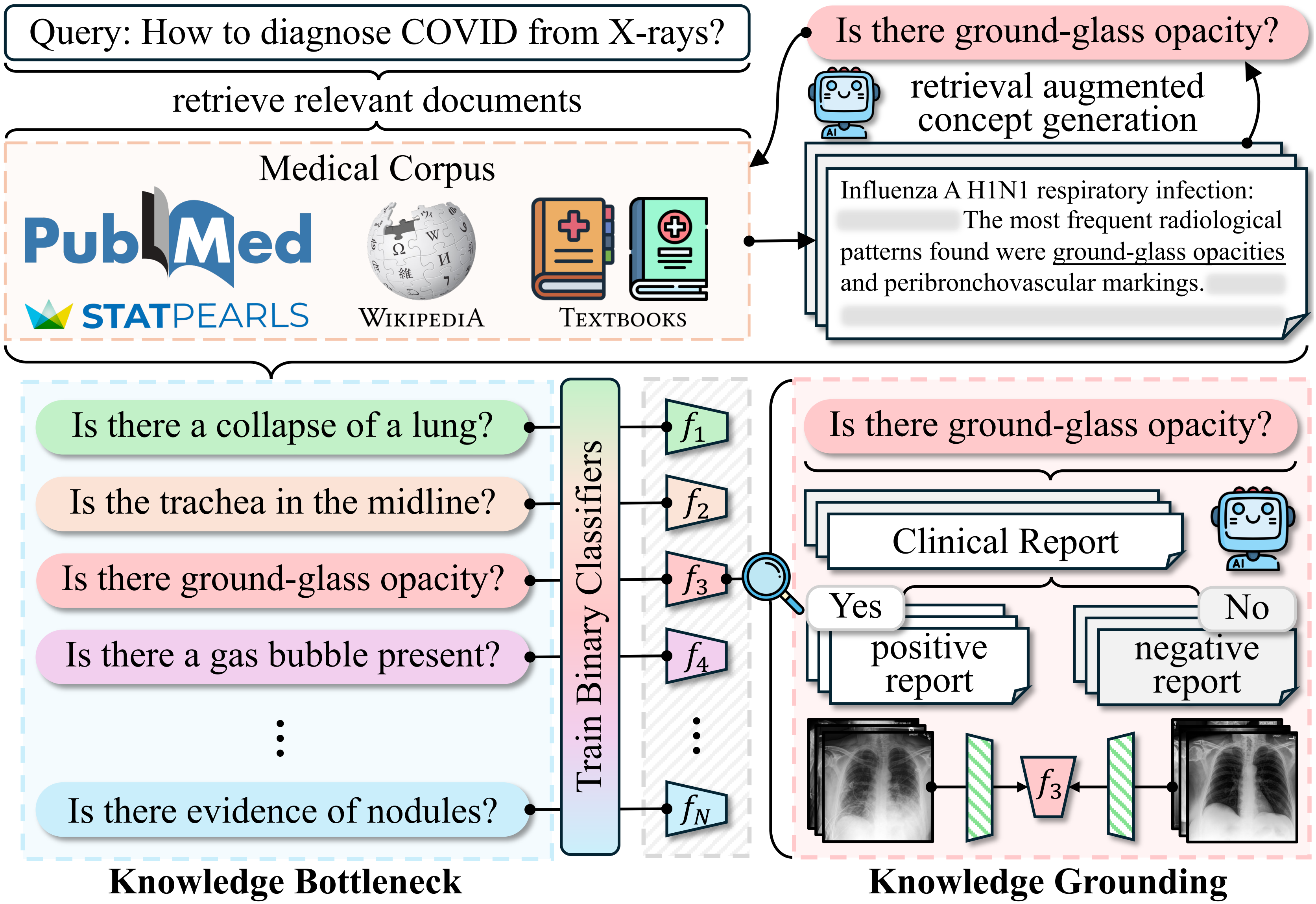
KnoBo System Overview



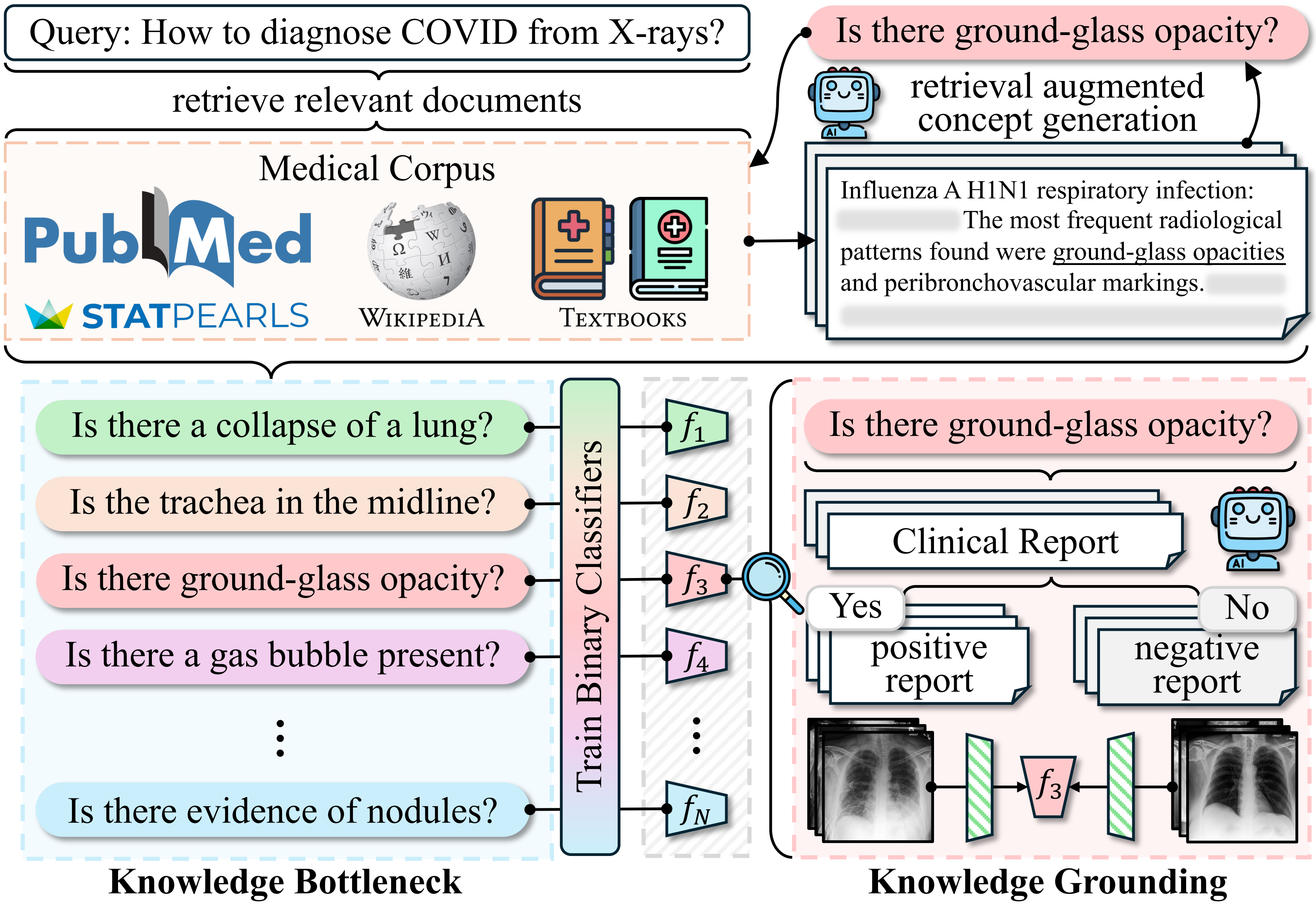
KnoBo System Overview



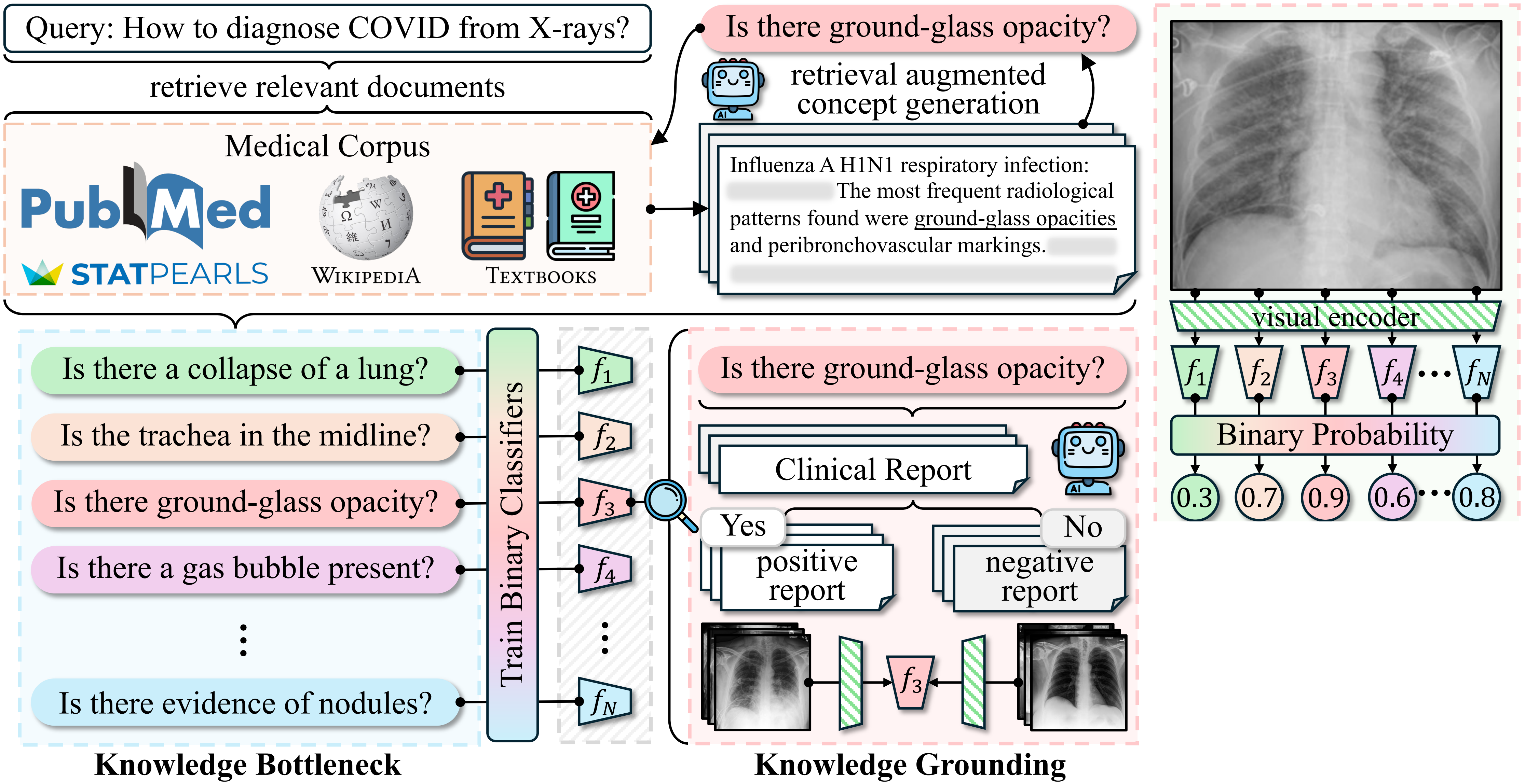
KnoBo System Overview



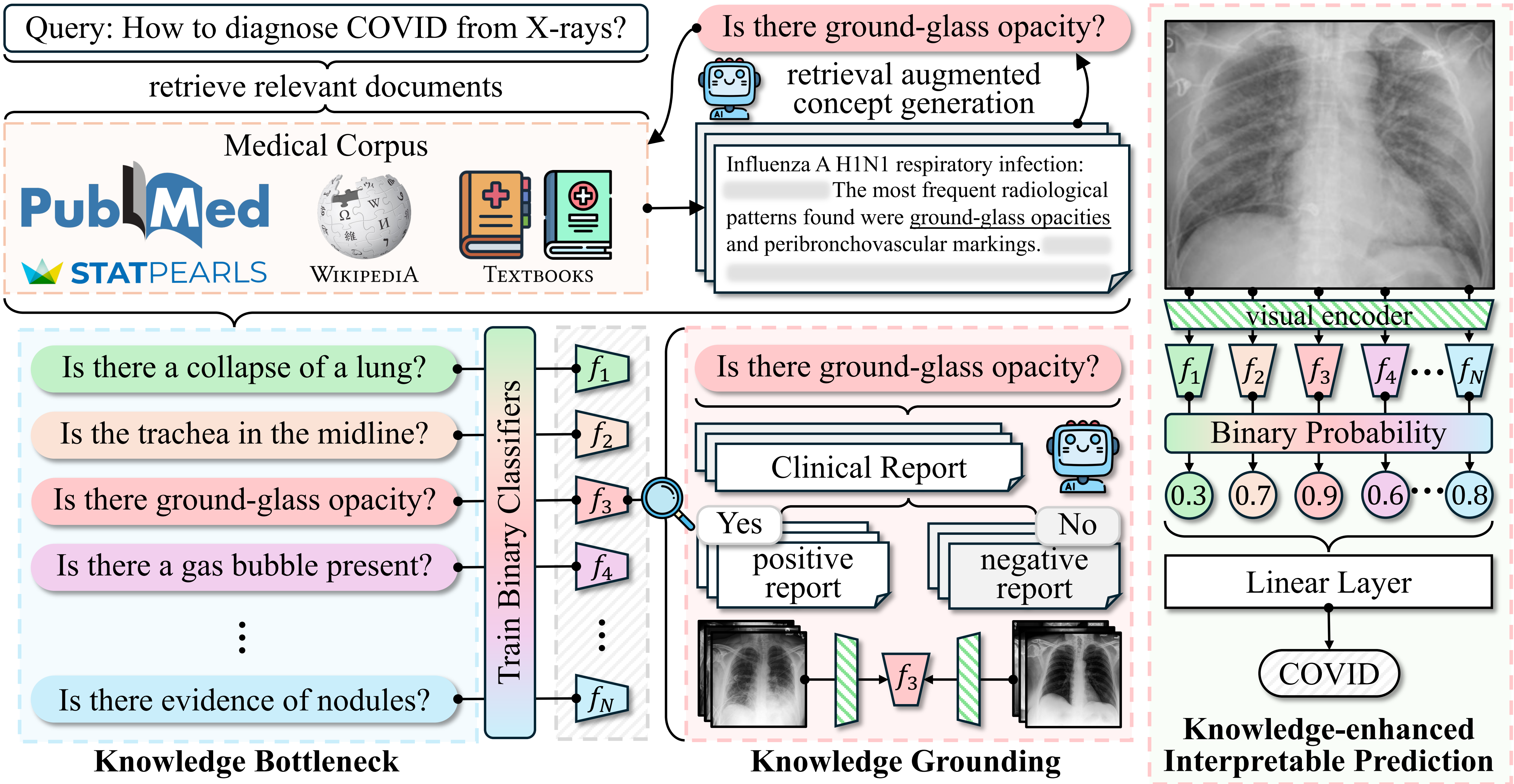
KnoBo System Overview



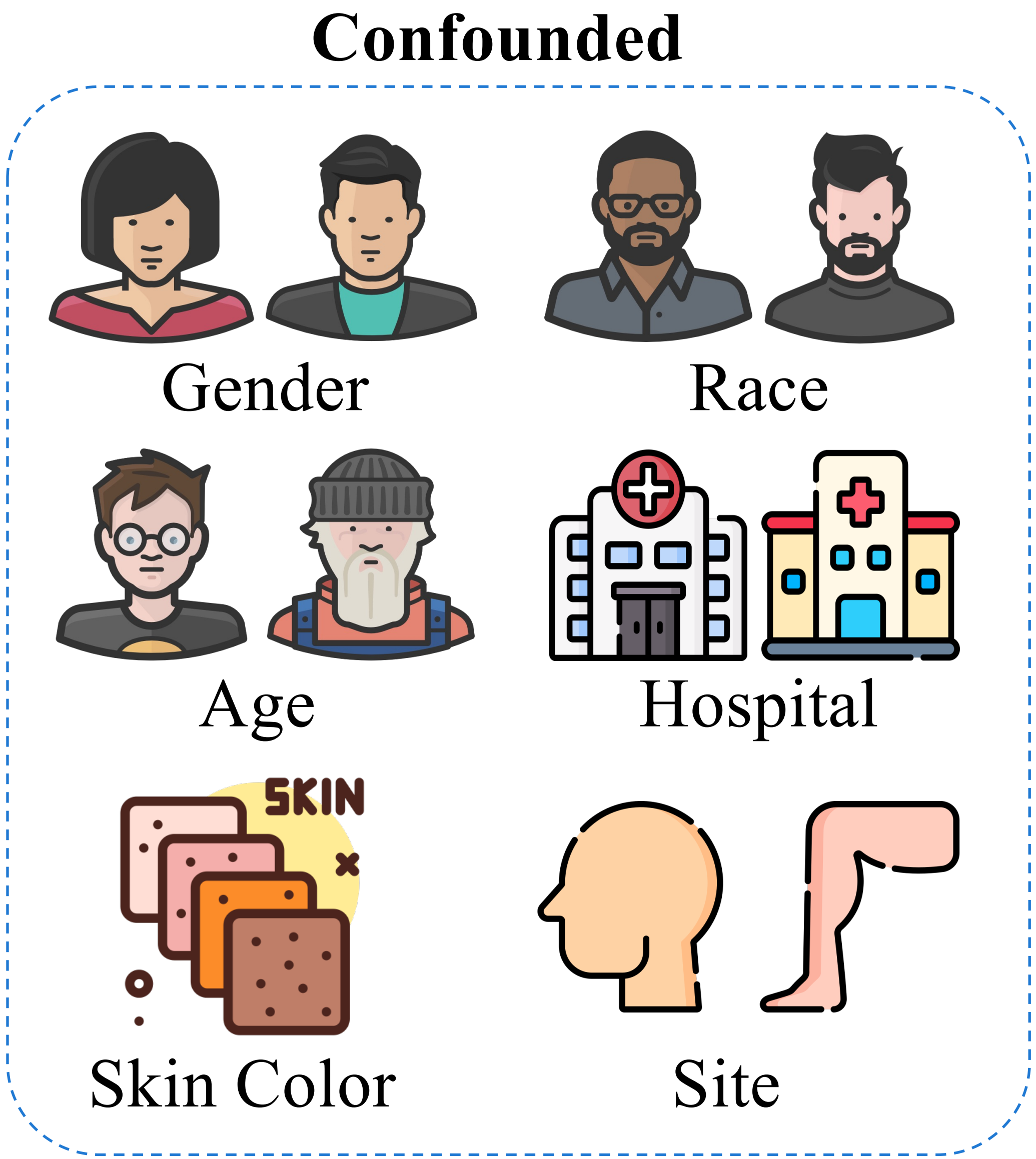
KnoBo System Overview

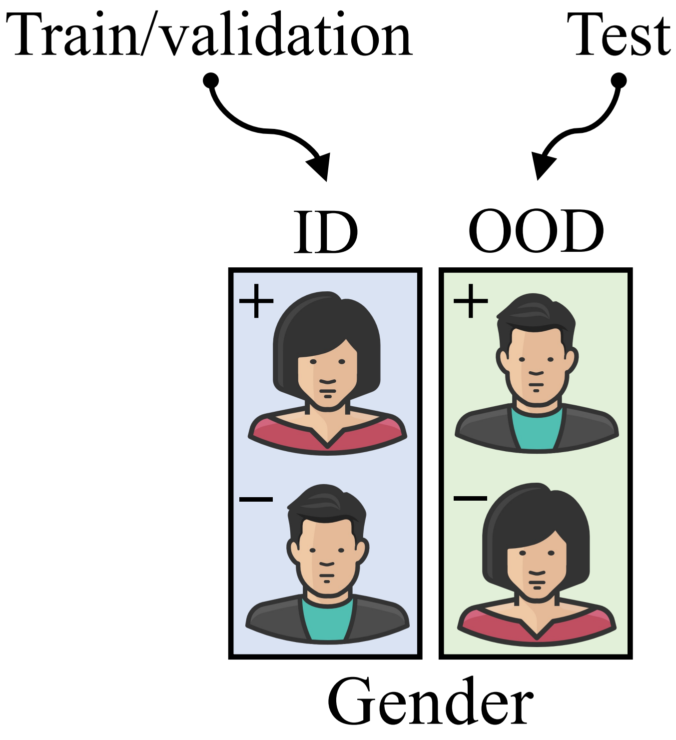


KnoBo System Overview



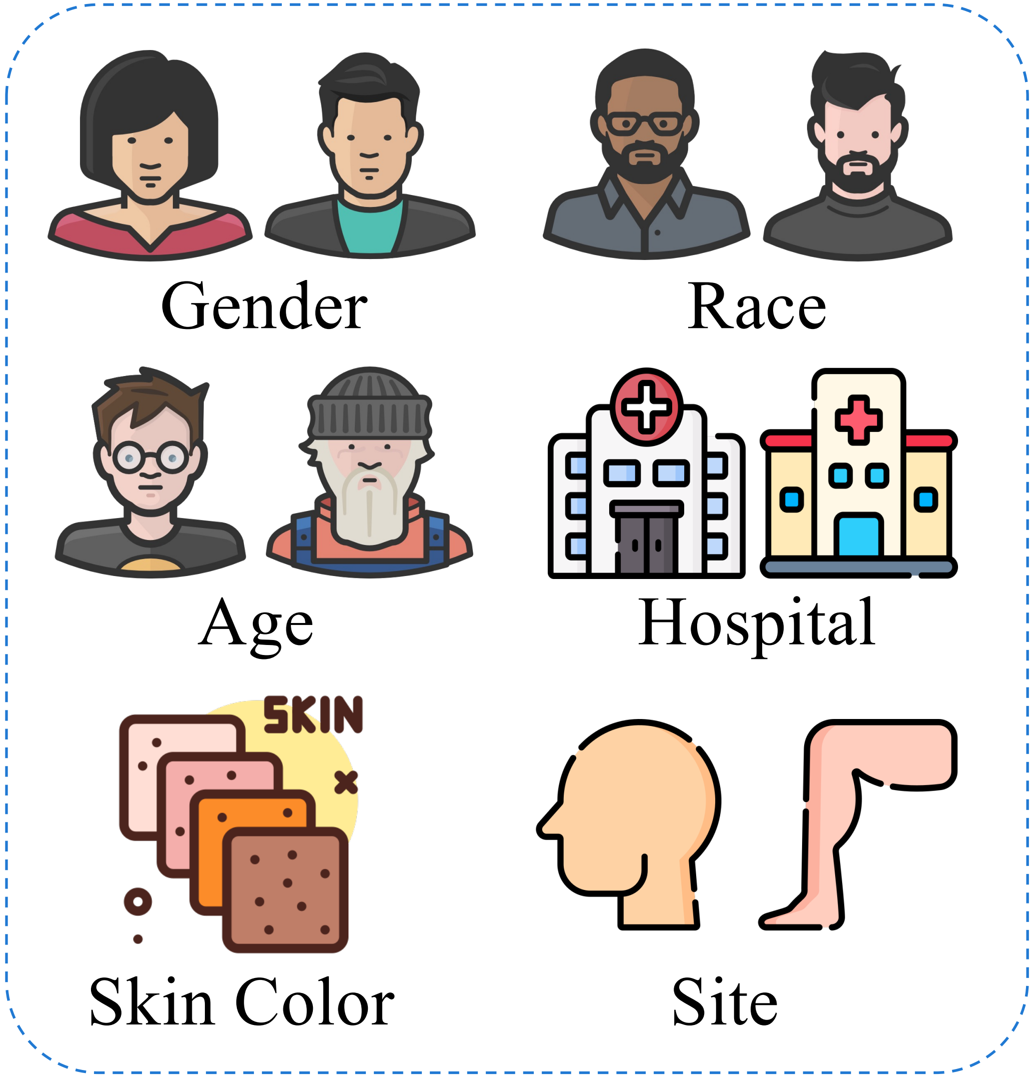
Datasets


Confounded
Gender
Race
Age
Hospital
SKIN
Skin Color
Site


Train/validation
Test
ID
OOD
+
−
+
−
Gender

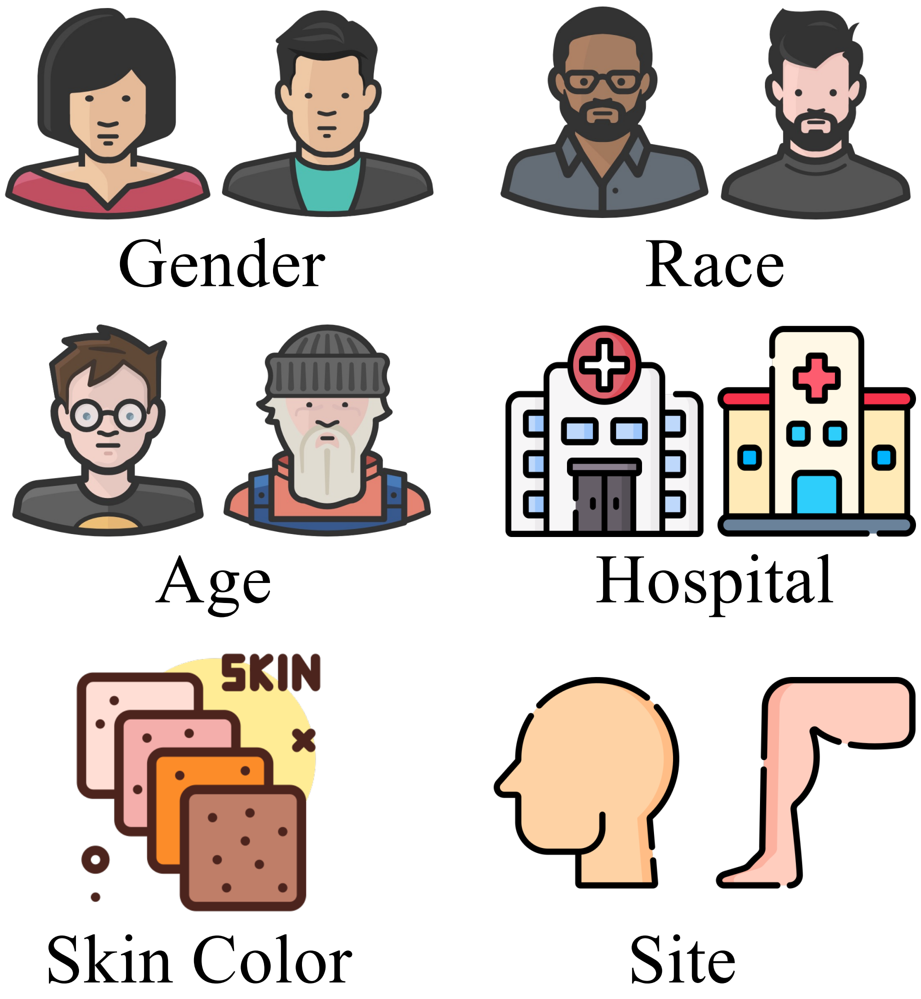
Datasets

Confounded



Datasets

Confounded



Standard

X-ray: Pneumonia, COVID-QU, NIH-CXR, Open-I, VinDr-CXR.

Skin Lesion: HAM10000, BCN20000, PAD-UFS-20, Melanoma, UWaterloo.

Experimental Setup

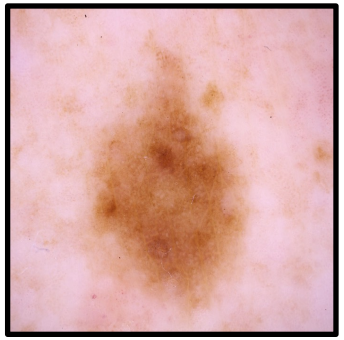
- **Baselines** (same vision backbone):
  - **Linear Probe**: logistic regression on the image features.
  - **End-to-end**: Unfreeze the visual encoder and update all parameters.
  - **LaBo**: knowledge priors from LLM, no knowledge grounding.
- **Metric**:
  - **Confounded datasets**: ID (validation), OOD (test), delta ↓ (|OOD-ID|), and domain-average accuracy (ID + OOD / 2).
  - **Standard datasets**: test accuracy.
  - **Overall Performance**: average over confounded and standard datasets.

Results on **X-ray** Datasets

Method	ID	OOD	delta↓	Domain Average	Standard	Overall
Linear Probe	95.2	30.7	64.5	62.9	**73.8**	68.4
End-to-End	**96.7**	17.0	79.7	56.8	70.2	63.5
LaBo	93.5	34.8	58.7	64.2	72.1	68.1
KnoBo	89.7	**58.8**	**30.9**	**74.3**	73.1	**73.7**

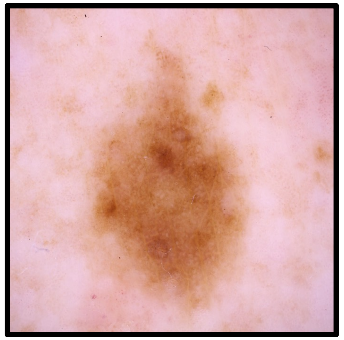
The best score is **bold** and the second best is underlined.

Results on **Skin Lesion** Datasets



Method	ID	OOD	delta↓	Domain Average	Standard	Overall
Linear Probe	91.9	52.1	39.8	72.0	82.8	77.4
End-to-End	**95.6**	47.6	48.0	71.6	**84.3**	77.9
LaBo	89.9	51.4	38.4	70.6	80.0	75.3
KnoBo	86.0	**70.5**	**14.1**	**78.3**	78.1	**78.2**

Results on **Skin Lesion** Datasets



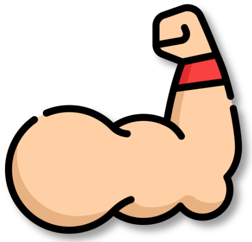
Method	**ID**	**OOD**	**delta↓**	**Domain Average**	**Standard**	**Overall**
Linear Probe	91.9	<u>52.1</u>	39.8	<u>72.0</u>	<u>82.8</u>	77.4
End-to-End	**95.6**	47.6	48.0	71.6	**84.3**	<u>77.9</u>
LaBo	89.9	51.4	<u>38.4</u>	70.6	80.0	75.3
KnoBo	86.0	**70.5**	**14.1**	**78.3**	78.1	**78.2**

KnoBo is more robust on confounded datasets.
KnoBo is competitive on standard datasets.

Comparison on Knowledge Types

Knowledge	X-ray Datasets			Skin Lesion Datasets		
	Confounded	Standard	Overall	Confounded	Standard	Overall
Prompt	72.9	72.8	<u>72.9</u>	**79.3**	72.8	76.0
Textbooks	72.0	<u>72.9</u>	72.4	<u>79.2</u>	76.4	77.8
Wikipedia	72.8	72.7	72.8	**79.3**	76.2	77.8
StatPearls	<u>73.4</u>	72.0	72.7	<u>79.2</u>	**77.6**	**78.4**
PubMed	**74.3**	**73.1**	**73.7**	**79.3**	<u>76.7</u>	<u>78.0</u>

Conclusion



Interpretable models with knowledge priors are **more robust in medical domains.**

Future Work

- Better feature representations for critical domains.
- Different structures of knowledge.
- Other usages of interpretable models.

Thank you!