

# Reliable and responsible data use

Pang Wei Koh  
University of Washington & AI2

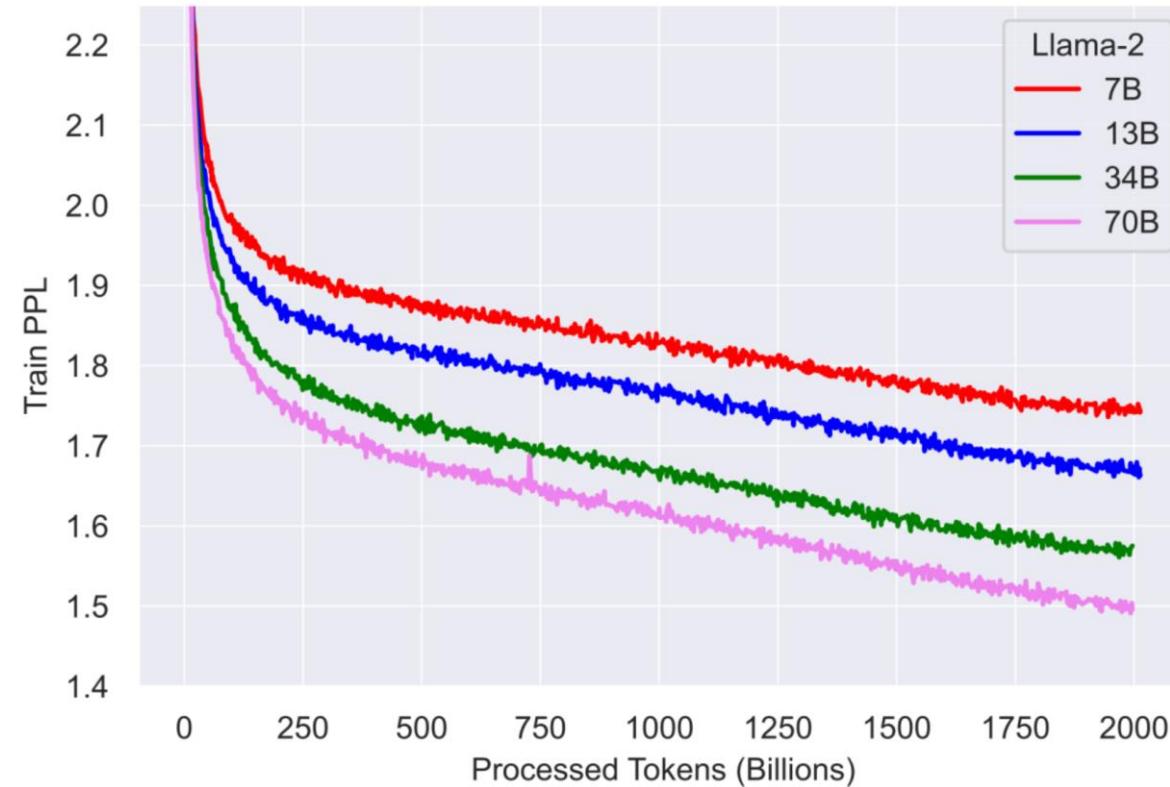


# Reliable and responsible data use

Pang Wei Koh  
University of Washington & AI2



# Data is driving progress in ML



# Data is driving progress in ML



5B image-text pairs

**nature**

Article | [Open access](#) | Published: 22 May 2024

**A whole-slide foundation model for digital pathology from real-world data**

1B histopathology tiles



3B protein sequences

# Data is driving progress in ML

- A lot of exciting work ahead on improving data quantity & quality
- 

**DataComp-LM: In search of the next generation of  
training sets for language models**

---



# This talk – beyond data sourcing & filtering

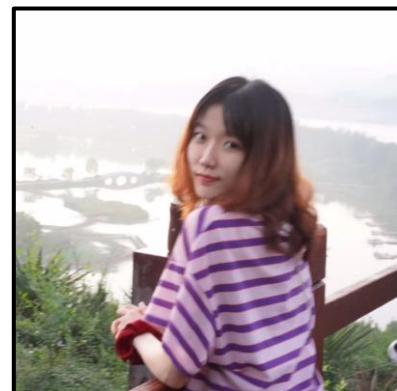
We'll cover:

1. Scaling up data at inference time
2. Training on synthetic data
3. Evaluating on interaction data

# Scaling up data at inference

**Scaling Retrieval-Based Language Models  
with a Trillion-Token Datastore**

Rulin Shao, Jacqueline He,  
Akari Asai, Weijia Shi, Tim Dettmers,  
Sewon Min, Luke Zettlemoyer, Pang Wei Koh. 2024.

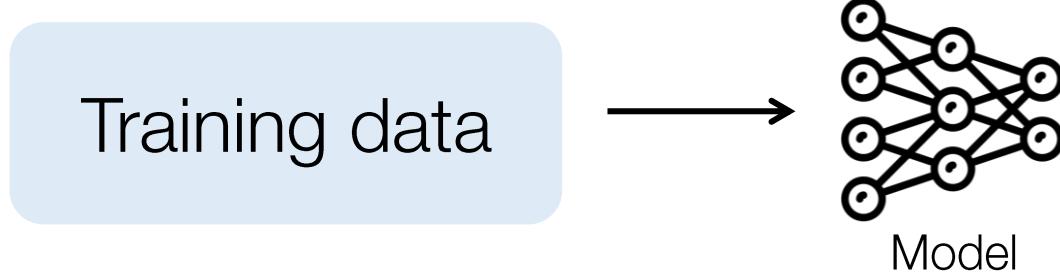


Rulin Shao

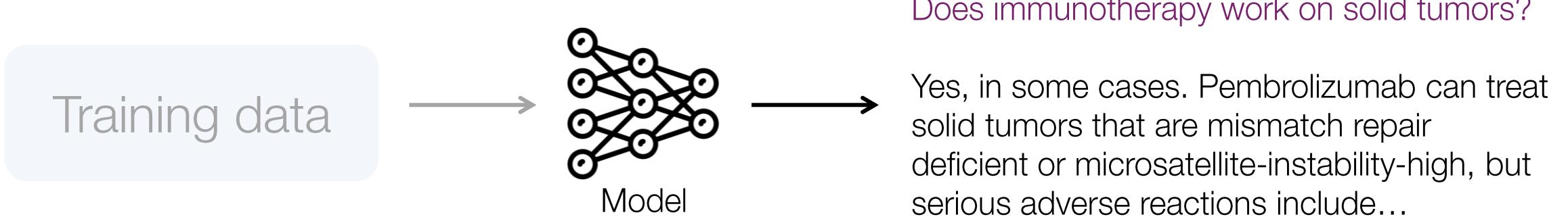


Jacqueline He

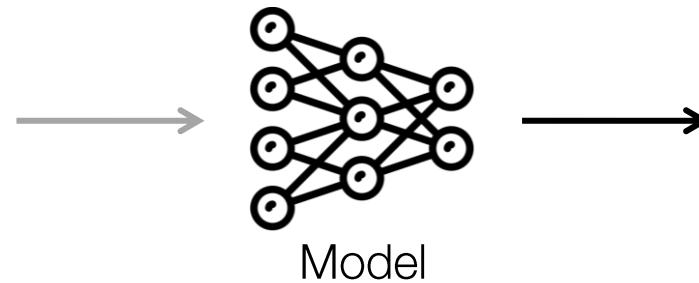
# Standard: Data stored in model parameters



# Standard: Data stored in model parameters

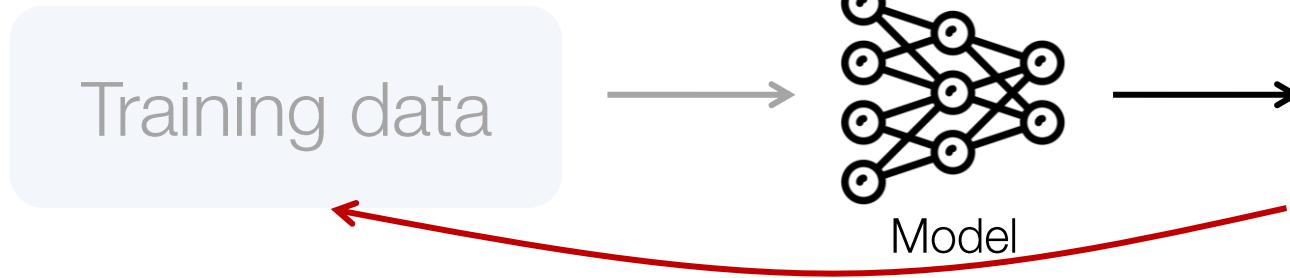


Training data



Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high, but serious adverse reactions include...

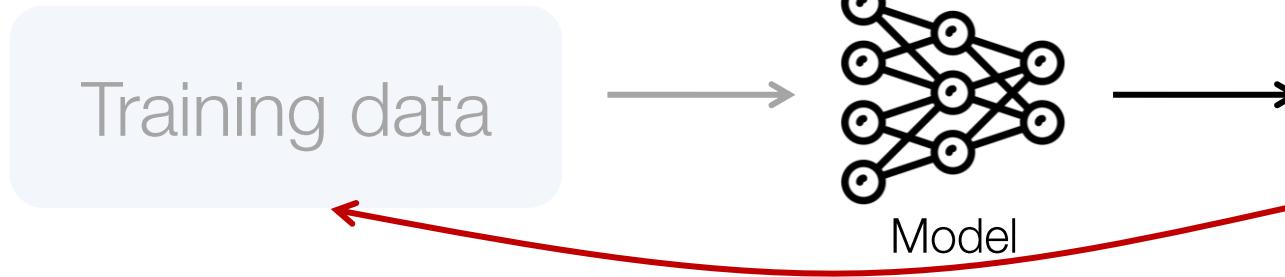


Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high [1], but serious adverse reactions include...

Doesn't naturally support:

1. Attribution for verification, credit
2. Using source info
3. Updating data



Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high [1], but serious adverse reactions include...

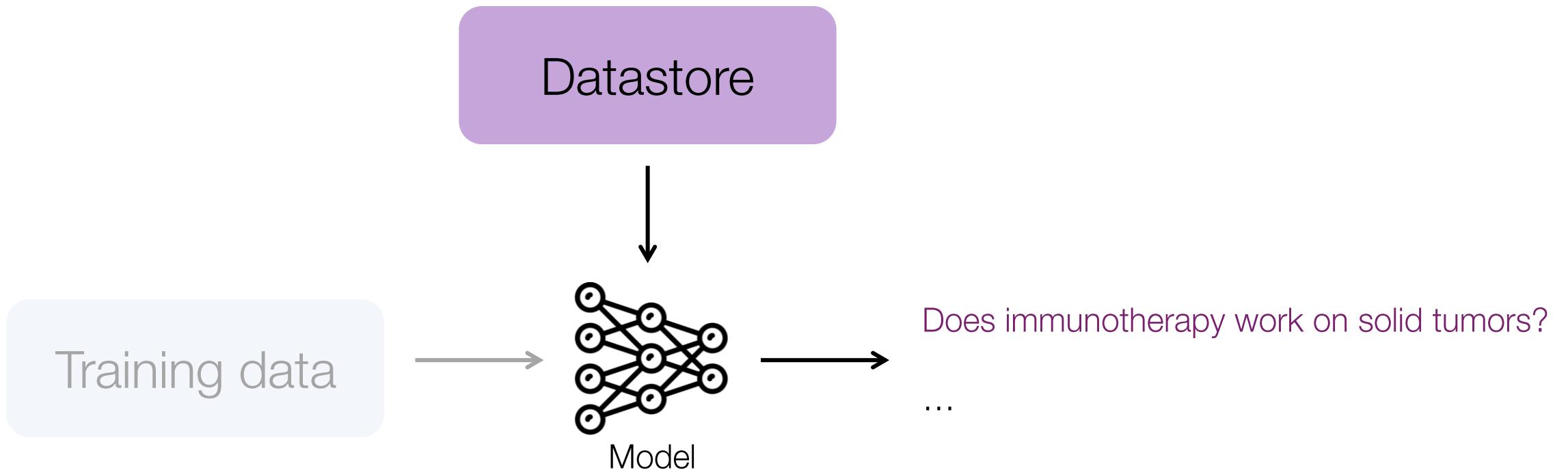
Doesn't naturally support:

1. Attribution for verification, credit
2. Using source info
3. Updating data

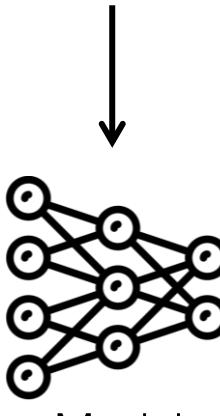
Mismatch between  
data vs. parameters

# Directly using data for inference via retrieval

- Can we build models where data points are first-class citizens?



## Datastore



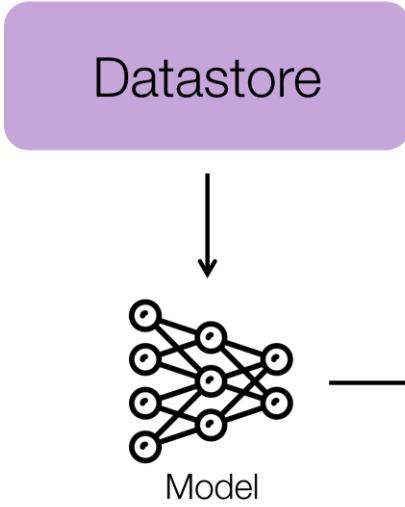
Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study  
Marabelle et al., Journal of Clinical Oncology, 2020

Genomes of tumors that are deficient in DNA mismatch repair (dMMR) have high microsatellite instability (MSI-H) and harbor hundreds to thousands of somatic mutations that encode potential neoantigens. Such tumors are therefore likely to be immunogenic...

Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high [1], but serious adverse reactions include...

[1] Marabelle et al., Results from KEYNOTE-158, Journal of Clinical Oncology, 2020.



Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study  
Marabelle et al., Journal of Clinical Oncology, 2020

Genomes of tumors that are deficient in DNA mismatch repair (dMMR) have high microsatellite instability (MSI-H) and harbor hundreds to thousands of somatic mutations that encode potential neoantigens. Such tumors are therefore likely to be immunogenic...

Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high [1], but serious adverse reactions include...

[1] Marabelle et al., Results from KEYNOTE-158, Journal of Clinical Oncology, 2020.

## Naturally supports:

1. Attribution
2. Using source info
3. Updating data

# Prior work on retrieval-based models

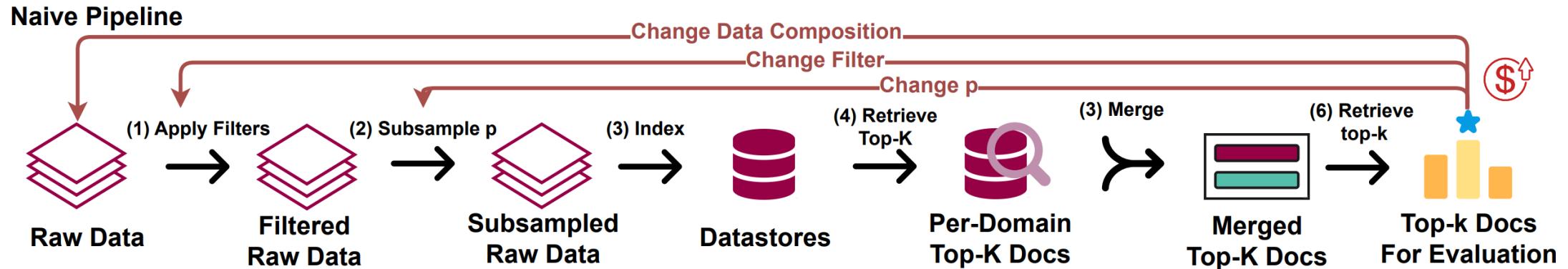
- Focus on small curated datastores, knowledge-intensive QA

Reference	# Tokens	Data Sources	Open sourced
ATLAS (Izacard et al., 2023)	<5B	Wikipedia	✗
REALM (Guu et al., 2020)	<5B	Wikipedia	✗
RALM (Ram et al., 2023)	<5B	Wikipedia	✓
SELF-RAG (Asai et al., 2024a)	<5B	Wikipedia	✓
REPLUG (Shi et al., 2023)*	47B	The Pile	✓
RA-DIT (Lin et al., 2024)	79B	Wikipedia, CommonCrawl	✗
SPHERE (Piktus et al., 2022)	90B	CCNet	✓
RETRO++ (Wang et al., 2024)	330B*	The Pile, CommonCrawl, RealNews, CC-Stories	✗
INSTRUCTRETRO (Wang et al., 2024)	1.2T*	Wikipedia, CommonCrawl, RealNews, CC-Stories, Books	✗
RETRO (Borgeaud et al., 2022)	1.7T*	MassiveText (Rae et al., 2022)	✗
MASSIVEDS (Ours)	<b>1.4T</b>	8 domains, listed in Table 2	✓

# Our work: Scaling the retrieval datastore

<b>Domain</b>	<b>Datasets</b>	<b>Size (B)</b>
BOOKS	RPJ Books	26.3
STEM	peS2o, RPJ ArXiv	97.7
ENCYCLOPEDIA	DPR 2018 Wiki, RPJ 2022 Wiki	31.9
FORUM (Q&A)	RPJ StackExchange	20.2
CODE	RPJ Github	52.8
MATH	OpenWebMath, NaturalProofs	14.1
BIOMEDICAL	PubMed	6.5
GENERAL WEB	RPJ CC (2019–2023), RPJ C4	1191.7
<b>Total</b>		<b>1441.2</b>

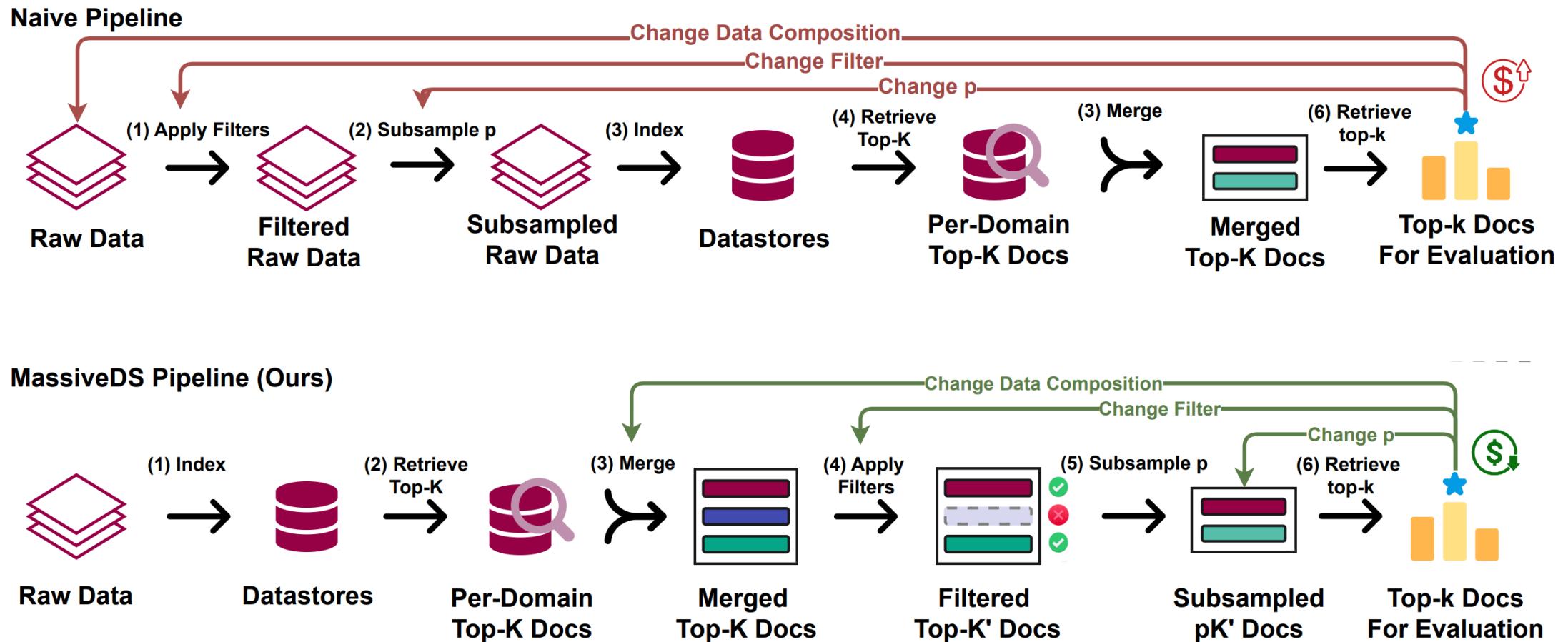
# Building datastores: Our pipeline



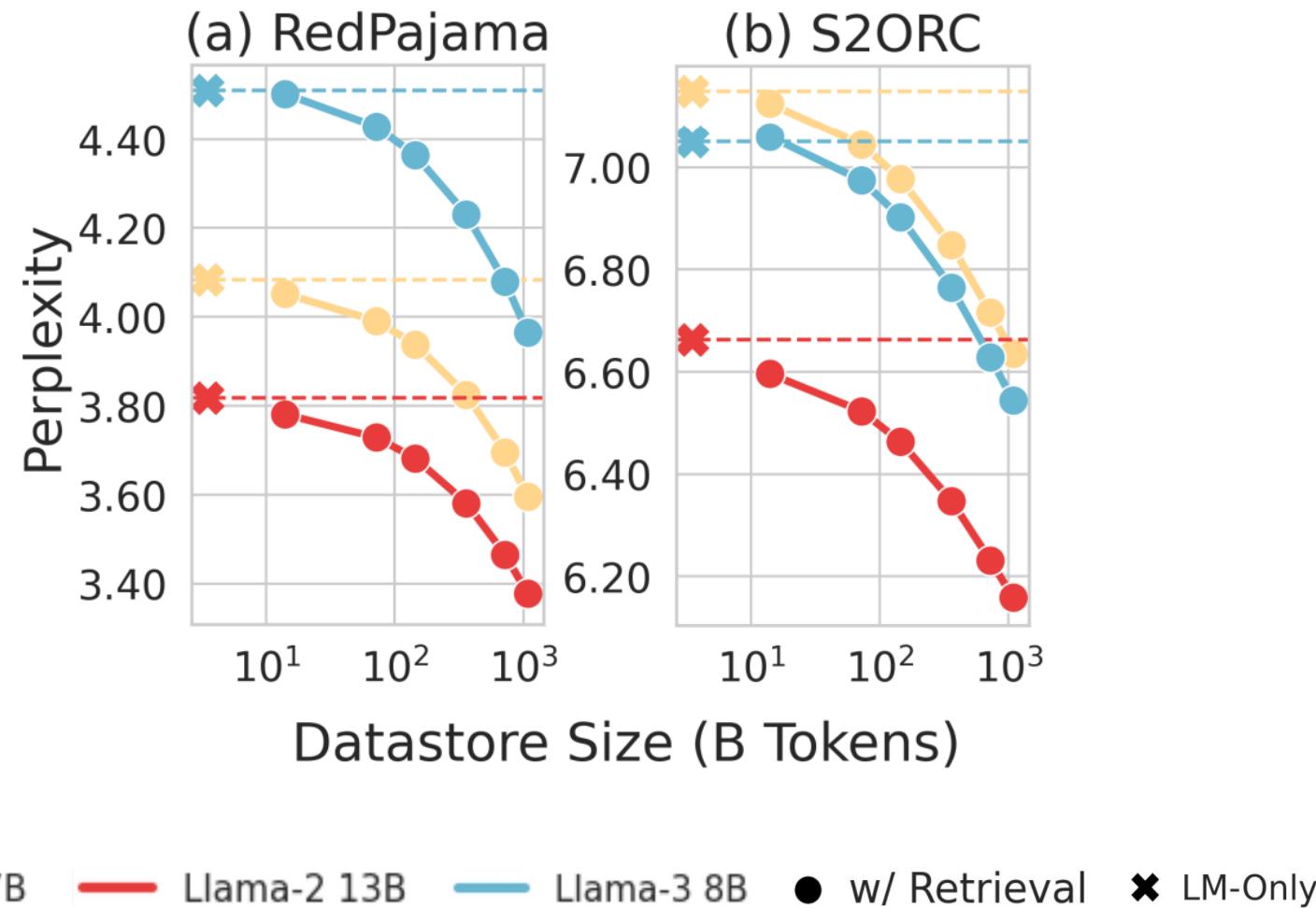
## Problem

Each different choice of size,  
composition, filters, etc. requires  
rebuilding the datastore

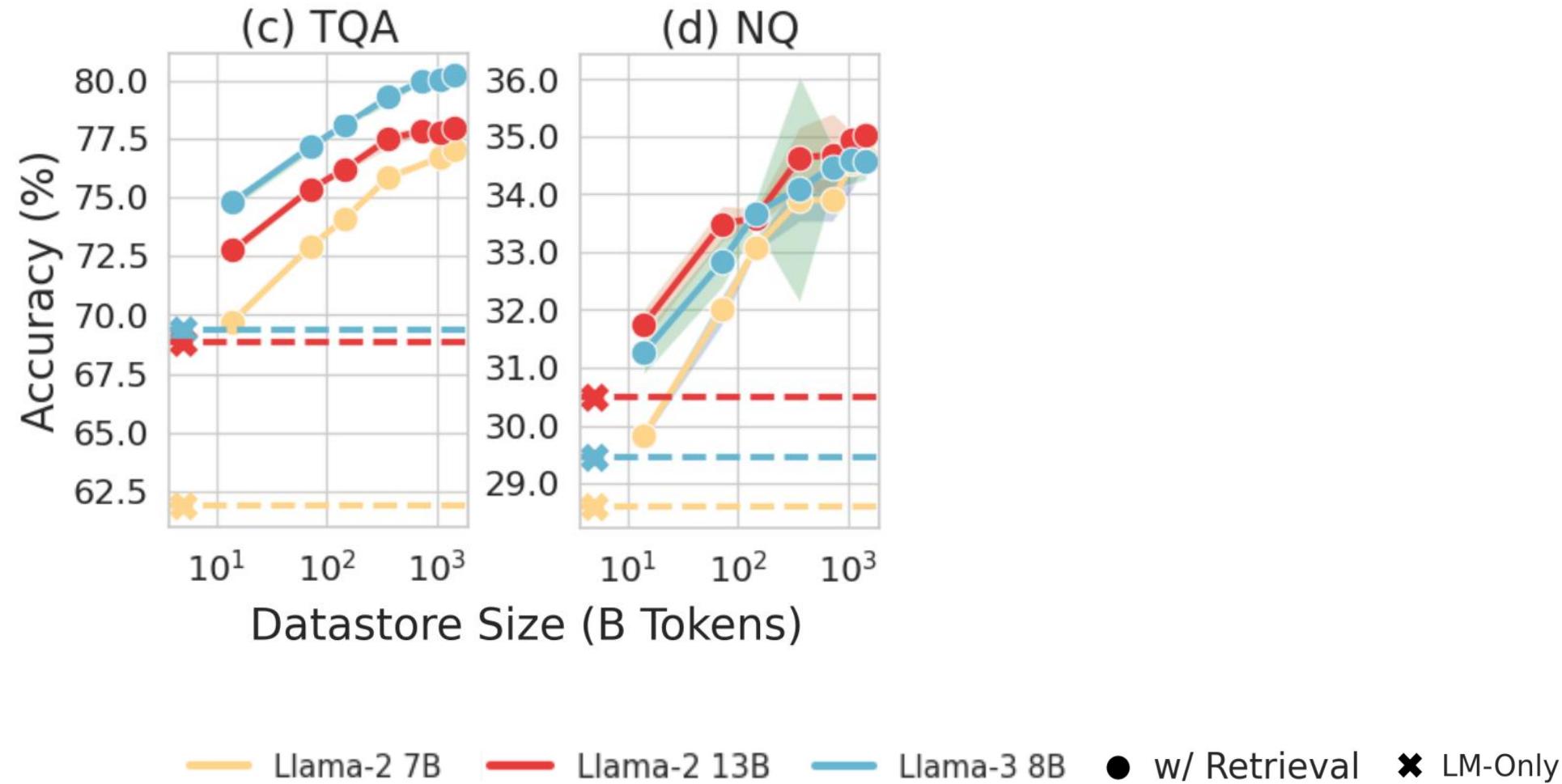
# Building datastores: Our pipeline



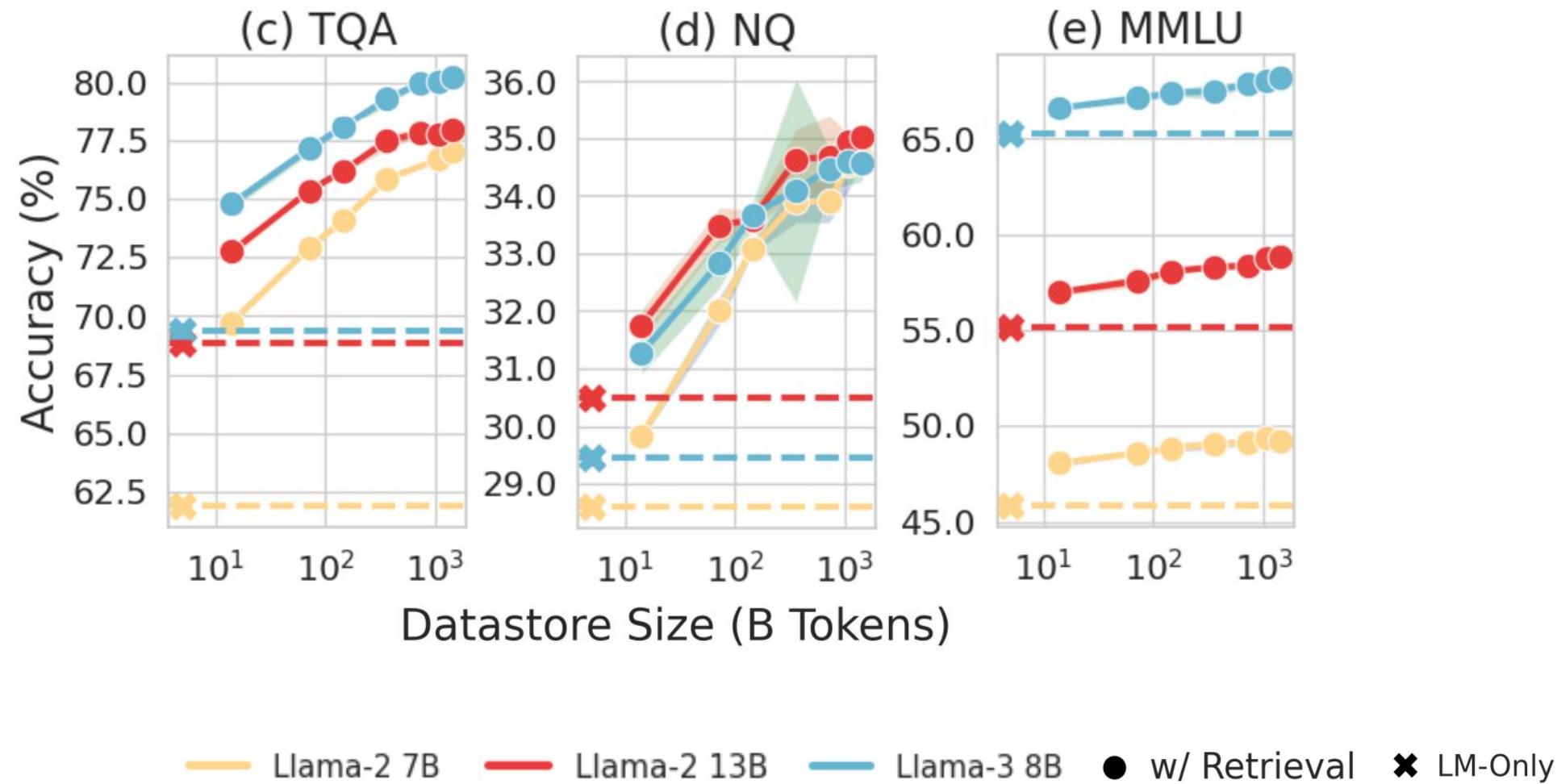
# DS scaling improves language modeling



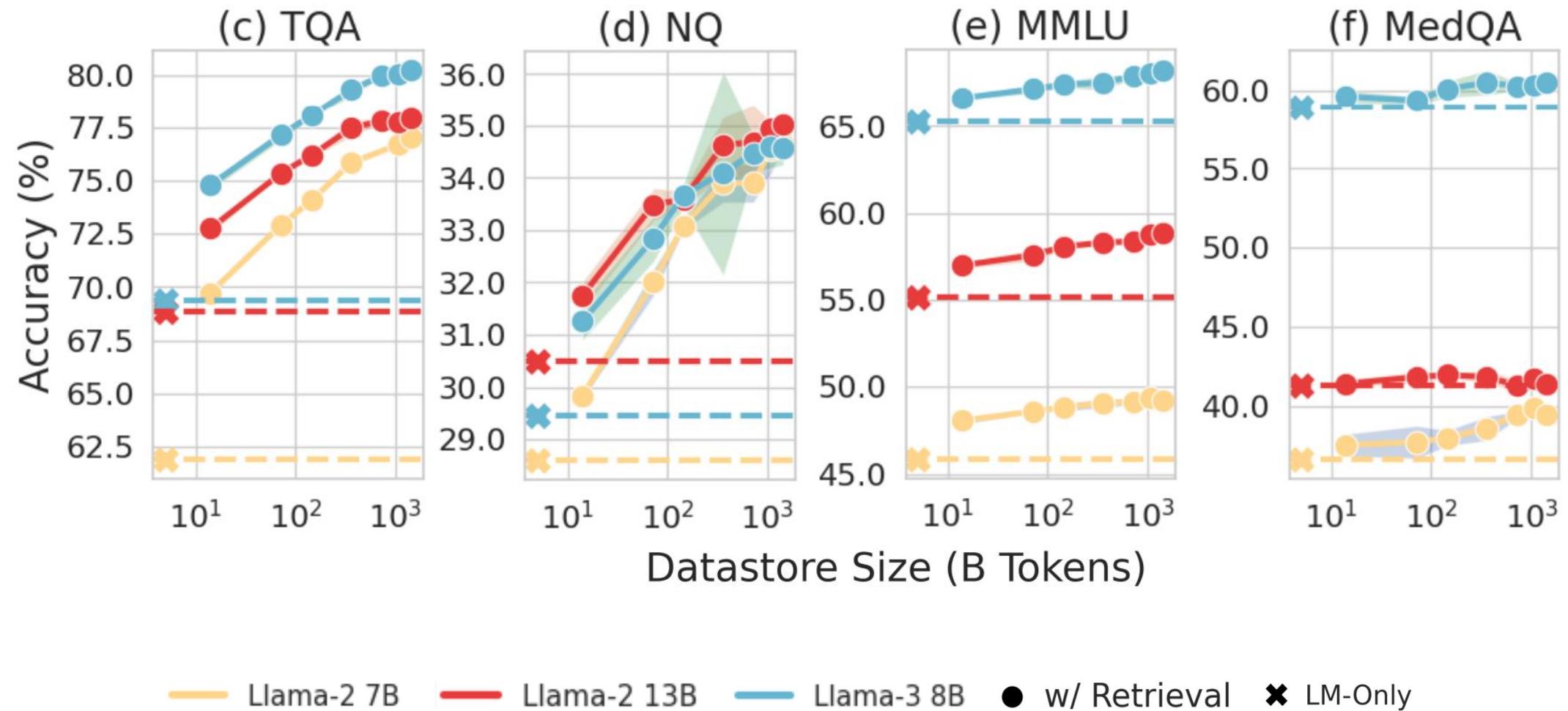
# DS scaling improves downstream tasks



# DS scaling improves downstream tasks



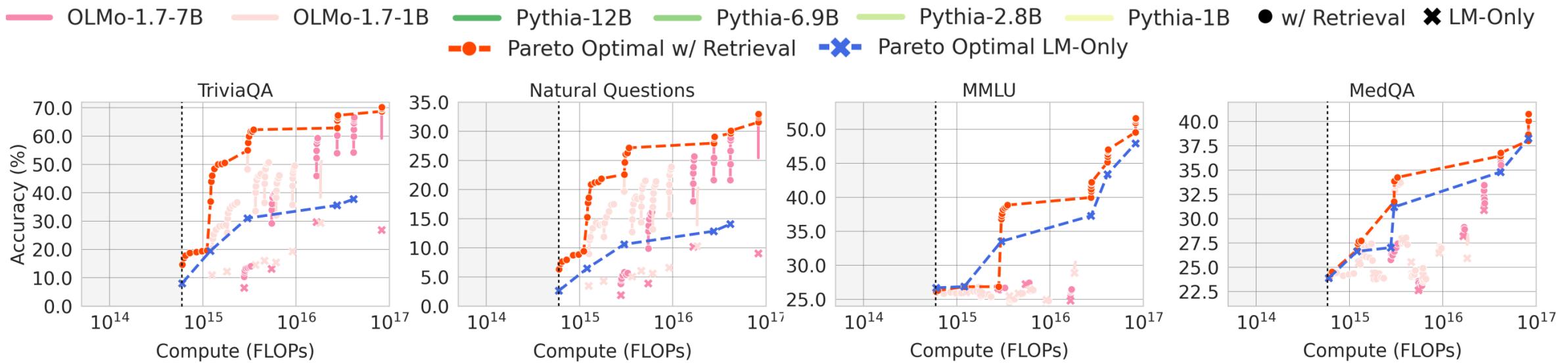
# DS scaling improves downstream tasks\*



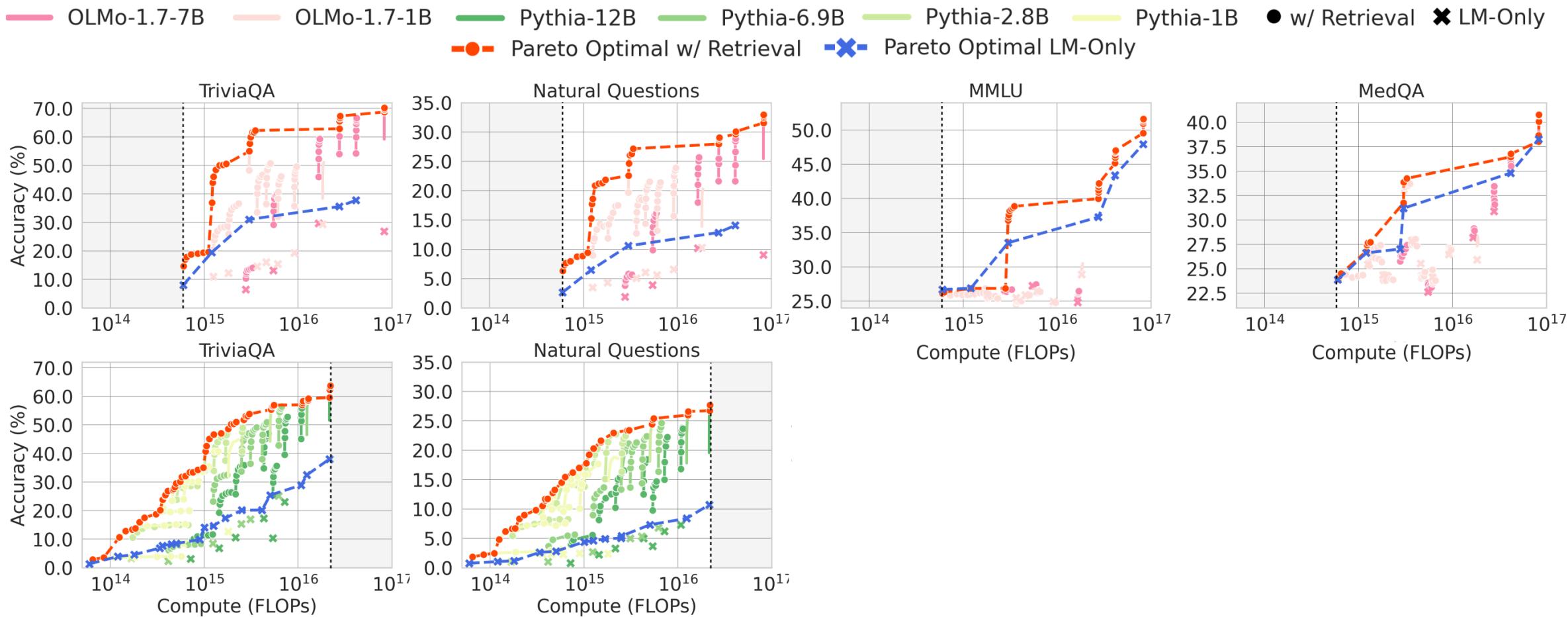
# Compute-optimal scaling

— OLMo-1.7-7B — OLMo-1.7-1B — Pythia-12B — Pythia-6.9B — Pythia-2.8B — Pythia-1B ● w/ Retrieval ✕ LM-Only  
—●— Pareto Optimal w/ Retrieval —✖— Pareto Optimal LM-Only

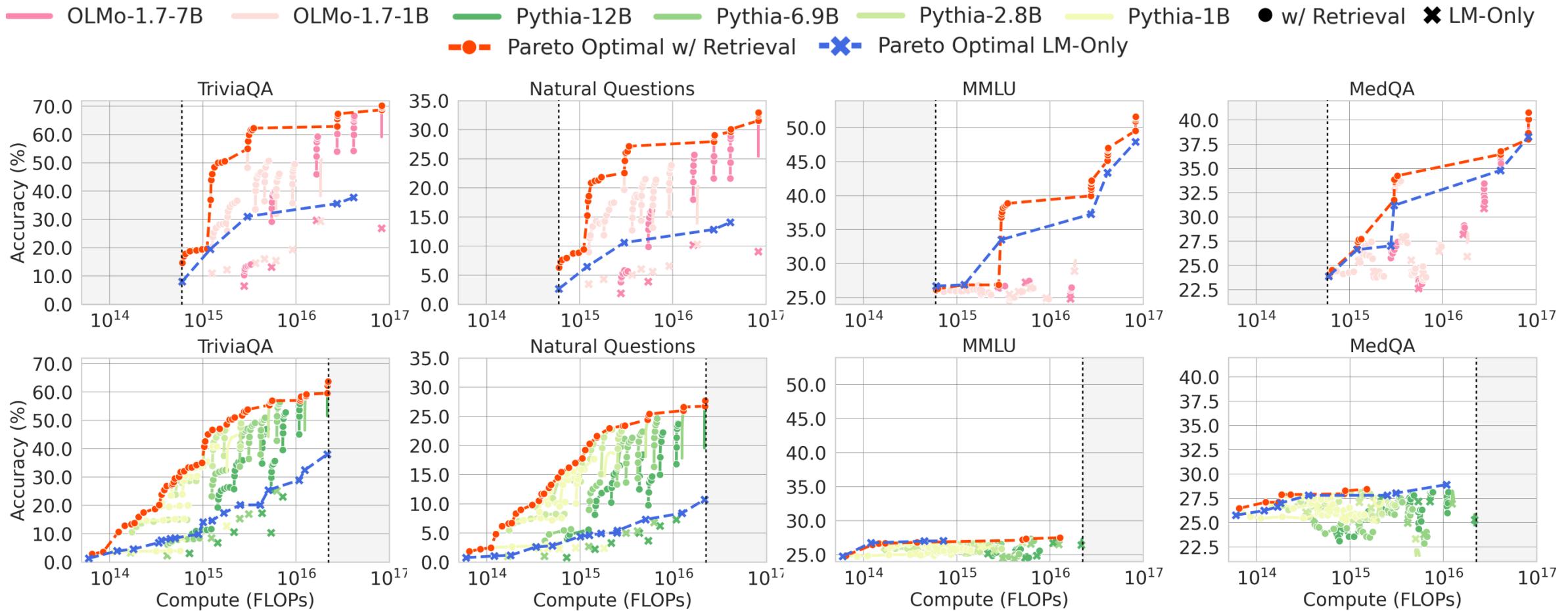
# Compute-optimal scaling



# Compute-optimal scaling



# Compute-optimal scaling

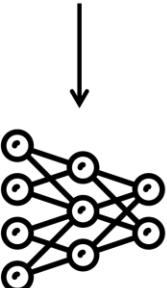


# Takeaways

- Scaling datastores in depth + breadth improves performance
- Efficient use of training-time and even inference-time compute
- Retrieval-based models allow us to reason about data as first-class citizens (while improving performance/efficiency)

# Future directions

Datastore



Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study

Marabelle et al., Journal of Clinical Oncology, 2020

Genomes of tumors that are deficient in DNA mismatch repair (dMMR) have high microsatellite instability (MSI-H) and harbor hundreds to thousands of somatic mutations that encode potential neoantigens. Such tumors are therefore likely to be immunogenic...

Does immunotherapy work on solid tumors?

Yes, in some cases. Pembrolizumab can treat solid tumors that are mismatch repair deficient or microsatellite-instability-high [1], but serious adverse reactions include...

[1] Marabelle et al., Results from KEYNOTE-158, Journal of Clinical Oncology, 2020.

1. Attribution
2. Using source info
3. Updating data

# Training on synthetic data

## **The Unmet Promise of Synthetic Training Images.**

Scott Geng, Cheng-Yu Hsieh, Vivek Ramanujan, Matthew Wallingford,  
Chun-Liang Li, Pang Wei Koh, Ranjay Krishna. 2024.



Scott Geng

# The promise of training on synthetic data

- Model distillation
  - Alpaca (Taori et al., 2023)
  - Phi-3 (Microsoft, 2024)

# The promise of training on synthetic data

- Model distillation
  - Alpaca (Taori et al., 2023)
  - Phi-3 (Microsoft, 2024)
- Can we use synthetic data to improve the best models?
  - Self-Instruct (Wang et al., 2022)
  - SynCLR (Tian et al., 2023)

**The emergent self-improvement capability.**

(Liu et al., 2024)

# Prior work: SynCLR

- Generate 150M captions (Llama2-7B) → 600M images (SD 1.5)

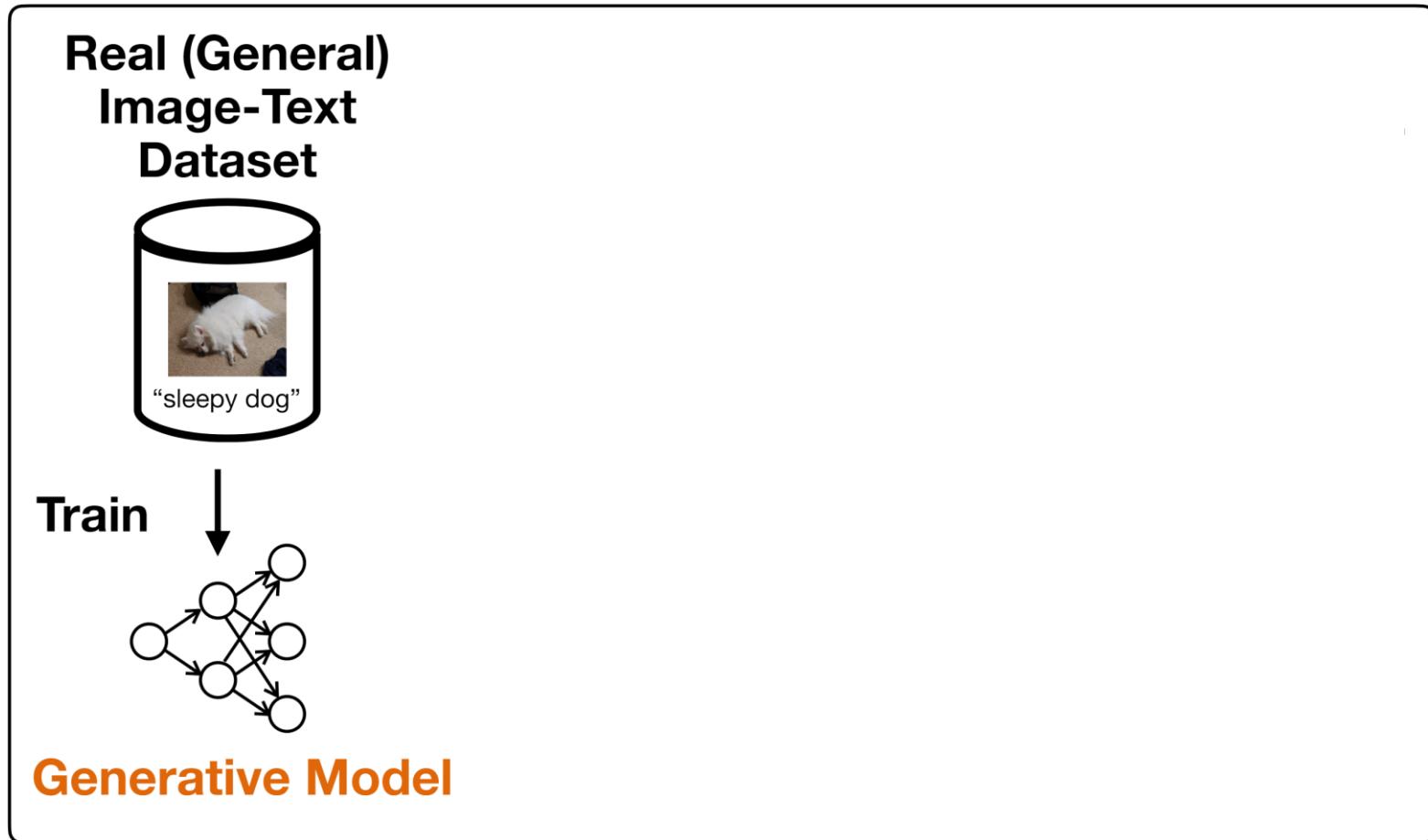
	text	img	# imgs		ImageNet	Aircraft	Cars	DTD	Flowers	Pets	SUN397	Caltech-101	Food-101	VOC2007	Average
CLIP	real	real	400M	ViT-B/16	80.2	59.5	86.7	79.2	98.1	93.1	78.4	94.7	92.8	89.2	85.7
				ViT-L/14	83.9	69.4	90.9	82.1	99.2	95.1	81.8	96.5	95.2	89.6	88.9
SynCLR	syn	syn	600M	ViT-B/16	80.7	81.7	93.8	79.9	99.1	93.6	76.2	95.3	91.6	89.4	<b>89.0</b>
				ViT-L/14	83.0	85.6	94.2	82.1	99.2	94.1	78.4	96.1	93.4	90.3	<b>90.4</b>

# Our work: Studying synthetic vs. real data

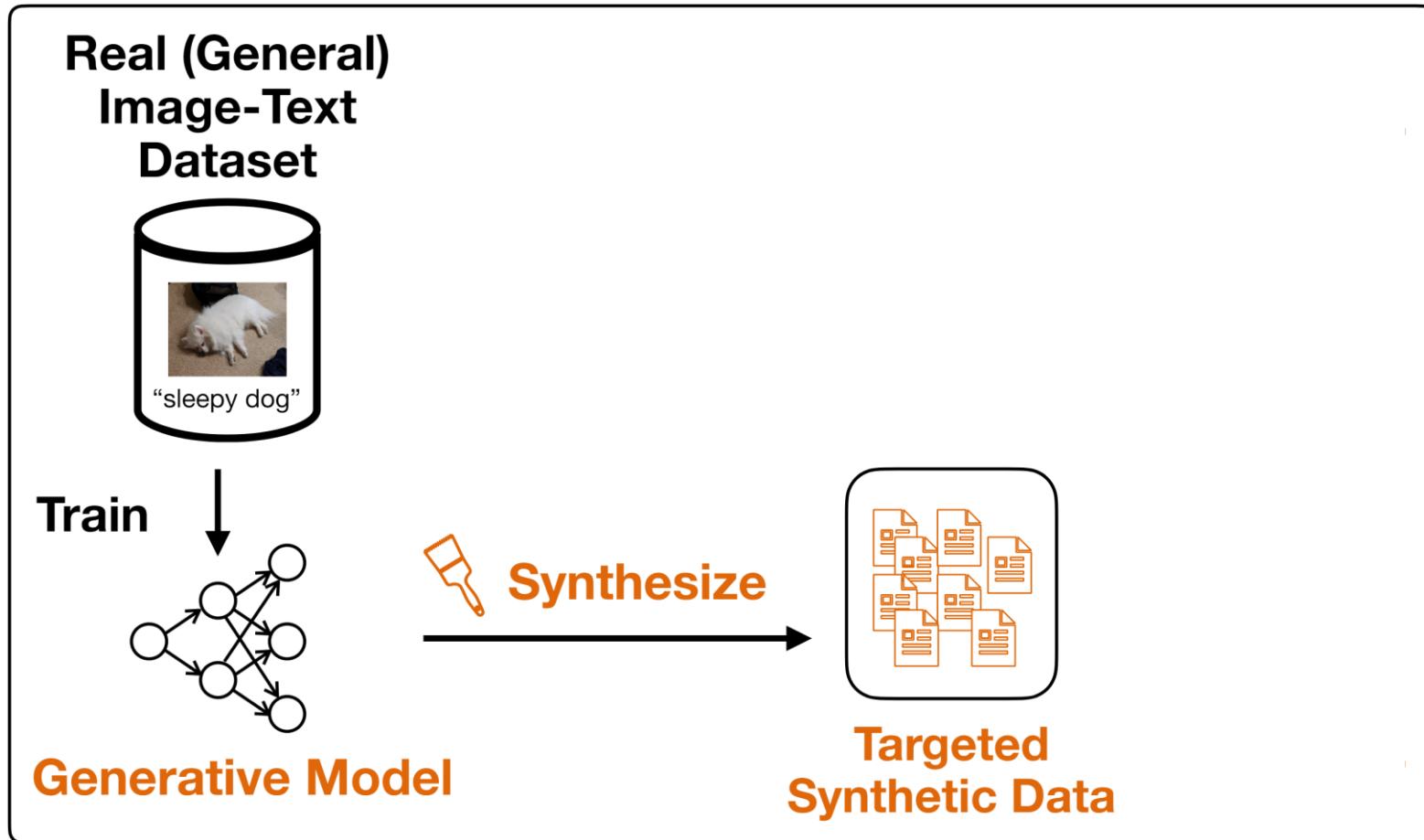
- What is driving these gains?
- Hypothesis: Controllability of the training distribution

source	prob.
IN-1k	0.47
Aircraft	0.05
Cars	0.05
Food	0.05
Flowers	0.03
Places-365, SUN397	0.09
IN-21k and others	0.26

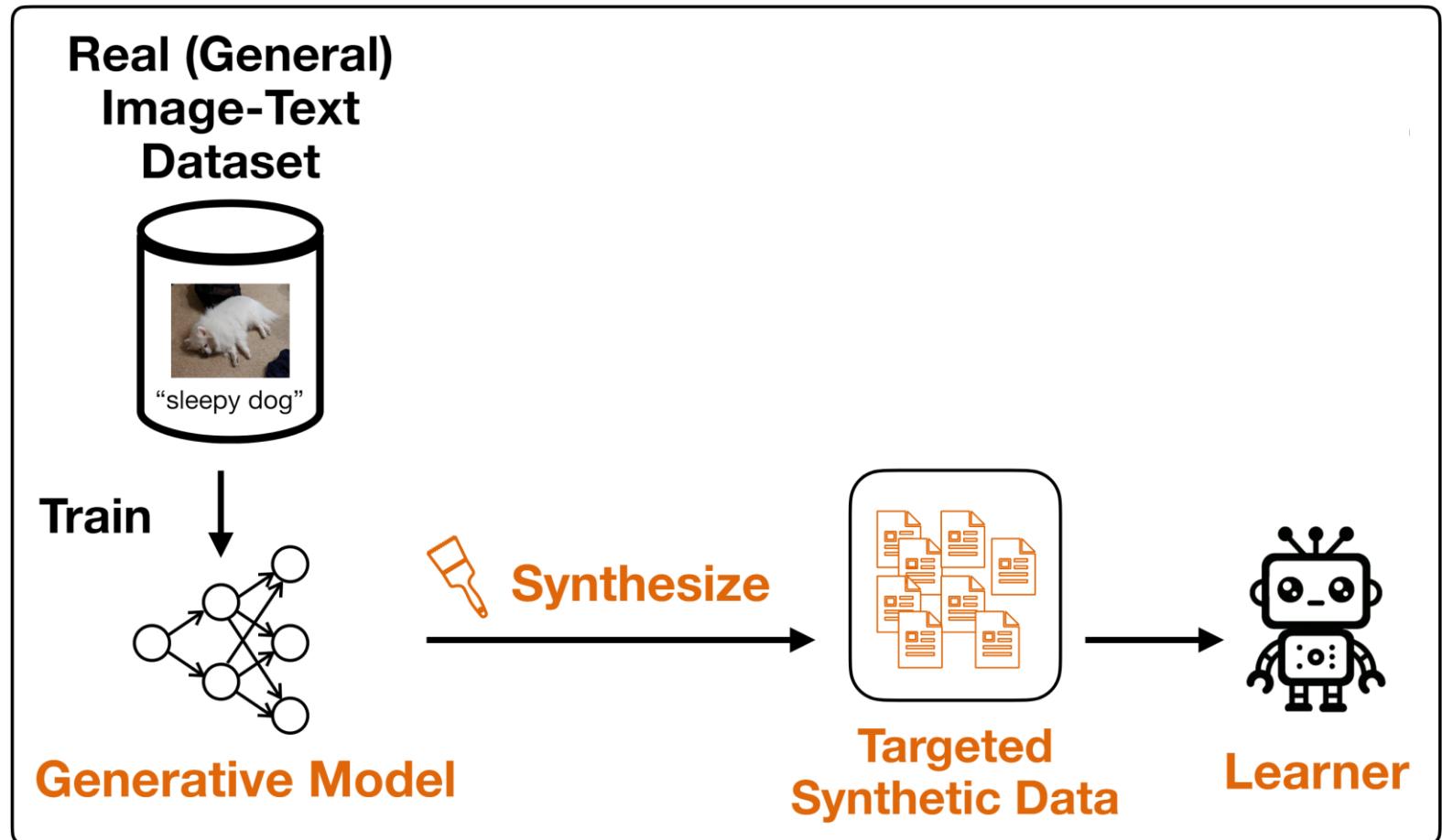
# Our work: Studying synthetic vs. real data



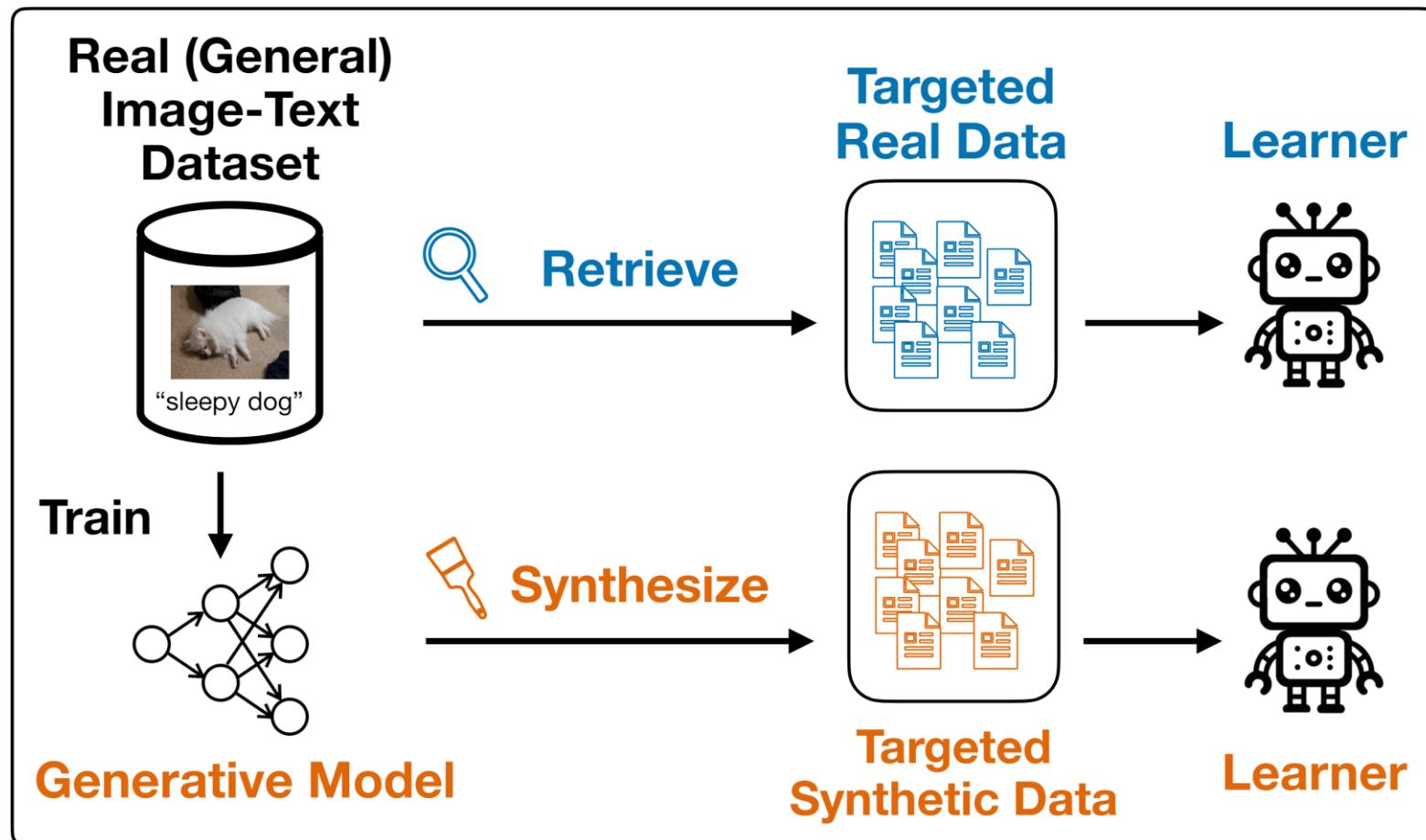
# Our work: Studying synthetic vs. real data



# Our work: Studying synthetic vs. real data



# Our work: Studying synthetic vs. real data



## Test Set Image

## Retrieved Real Images

## Generated Synthetic Images

“Airbus A320”



“Cessna 172”

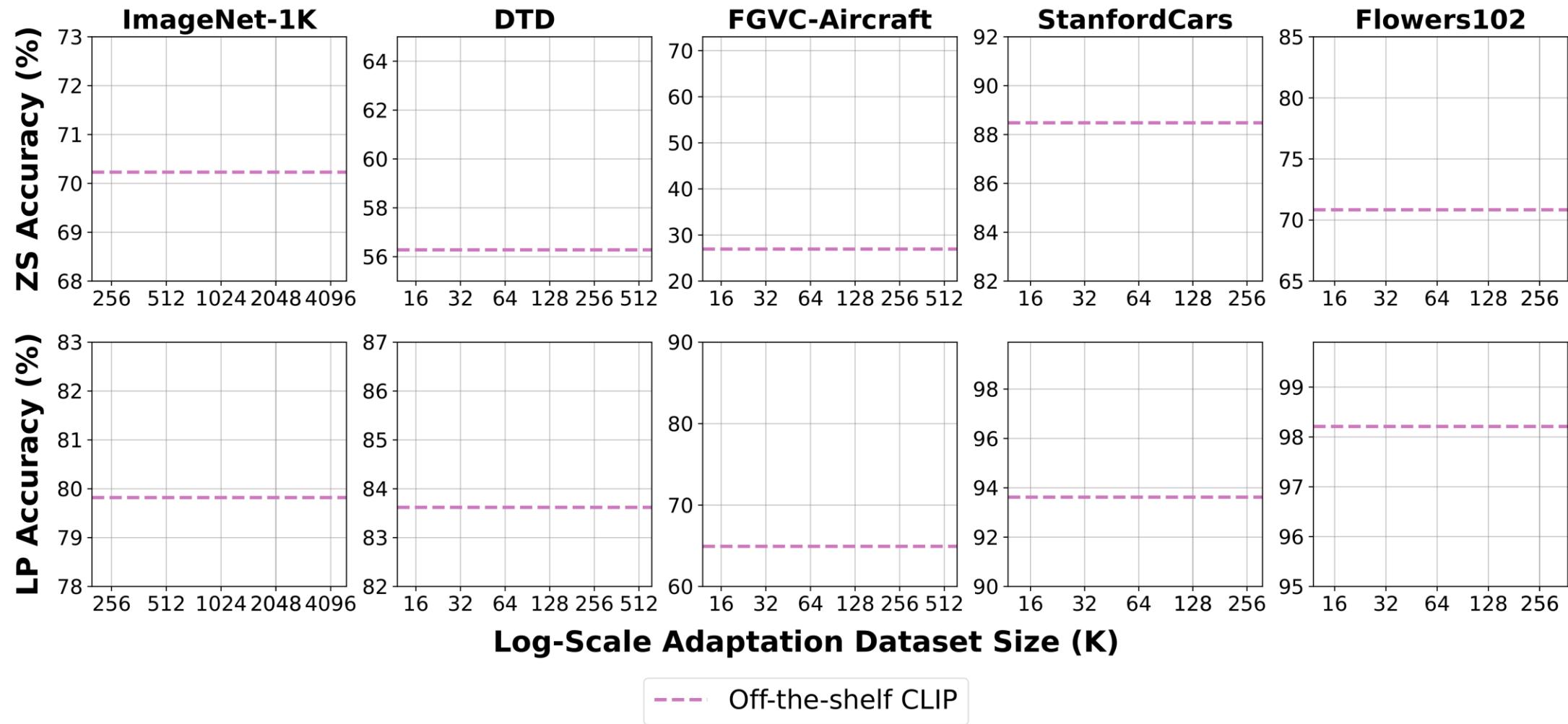


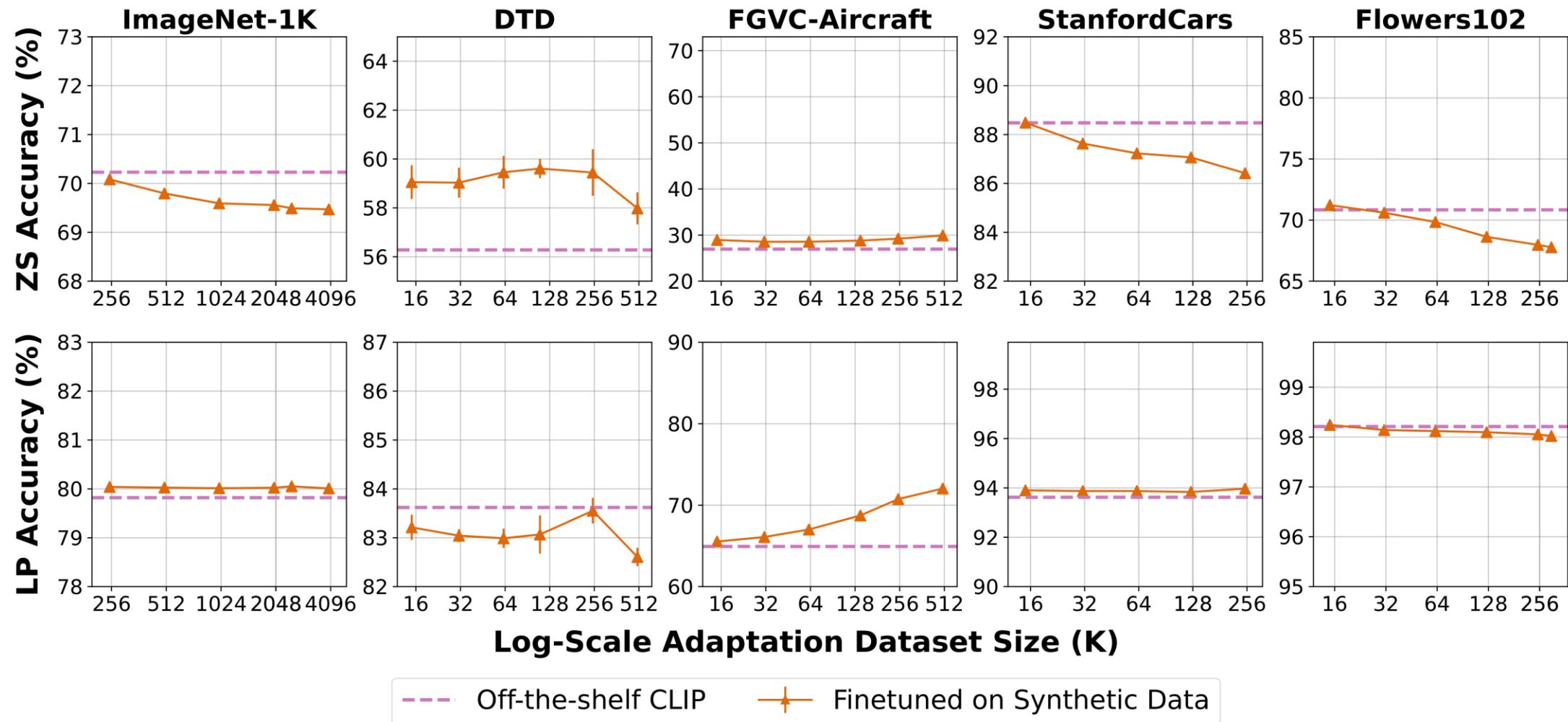
“Flute”



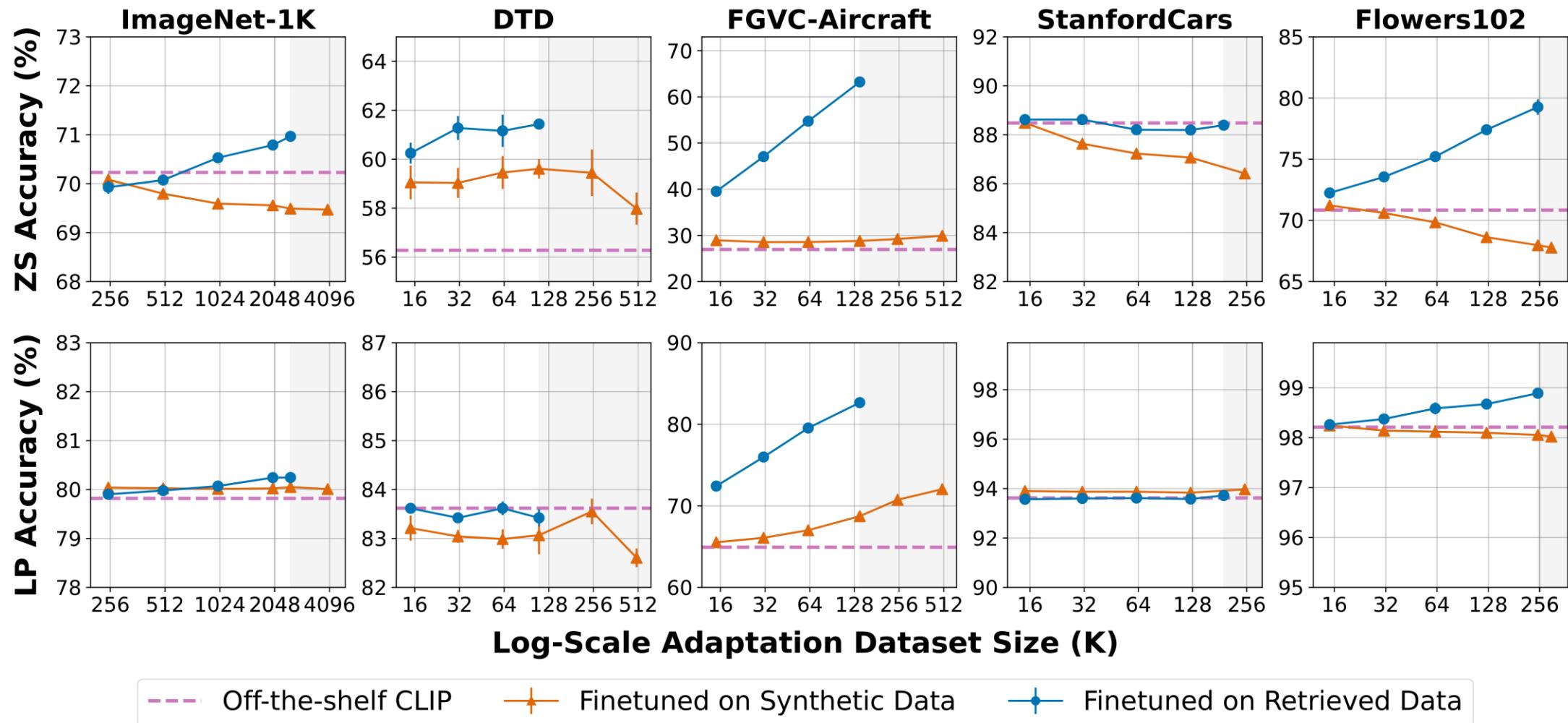
“Tabby Cat”







- Synthetic data helps\*



- Synthetic data helps\*
- Retrieved data > synthetic data

# Issues with synthetic data

- General distortions



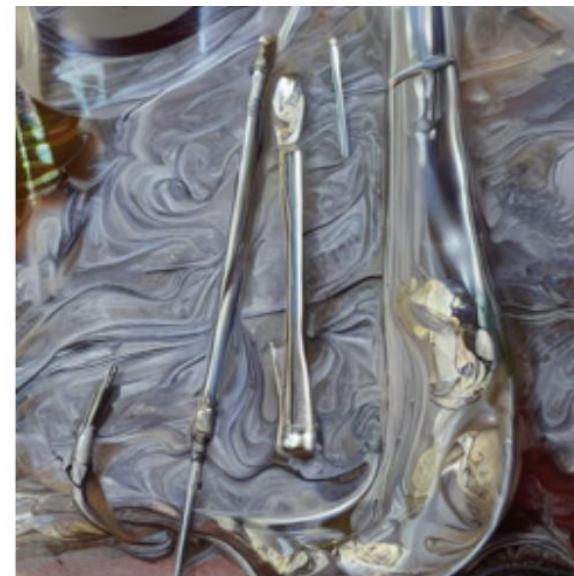
# Issues with synthetic data

- General distortions
- Class-relevant distortions



# Issues with synthetic data

- General distortions
- Class-relevant distortions
- Can be completely wrong



# Issues with synthetic data

- General distortions
- Class-relevant distortions
- Can be completely wrong
- Differences in “semantic” distribution

**“Tabby Cat”**



# Takeaways

- Synthetic data allows controllability (e.g., sampling more images from target classes)
- For image classification, currently: synthetic data < retrieved data
  - Burg et al., 2023; Zhou et al., 2023
- Definitely not ruling out synthetic data
- Open question: When and why should we use synthetic data?

# 1) Synthetic data + external models



**Multilingual Diversity Improves Vision-Language Representations**  
Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma,  
Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, Ranjay Krishna. 2024.



Thao Nguyen

# 2) Synthetic data + verifiers

## 1. Information-maximizing objective for Summarization

Document  $x$  - Summary  $y$

" **Find  $(x,y)$  such that  $PMI(x; y) > \alpha$  and  $\frac{|y|}{|x|} < \beta$**  "

**Information-Theoretic Distillation for Reference-Less Summarization**  
Jaehun Jung, Ximing Lu, Liwei Jiang, Faeze Brahman, Peter West,  
Pang Wei Koh, Yejin Choi. 2024.



Jaehun Jung

# Evaluating on interaction data

**MEDIQ: Question-Asking LLMs for Adaptive and Reliable Clinical Reasoning**

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen,  
Emma Pierson, Pang Wei Koh, Yulia Tsvetkov. 2024.

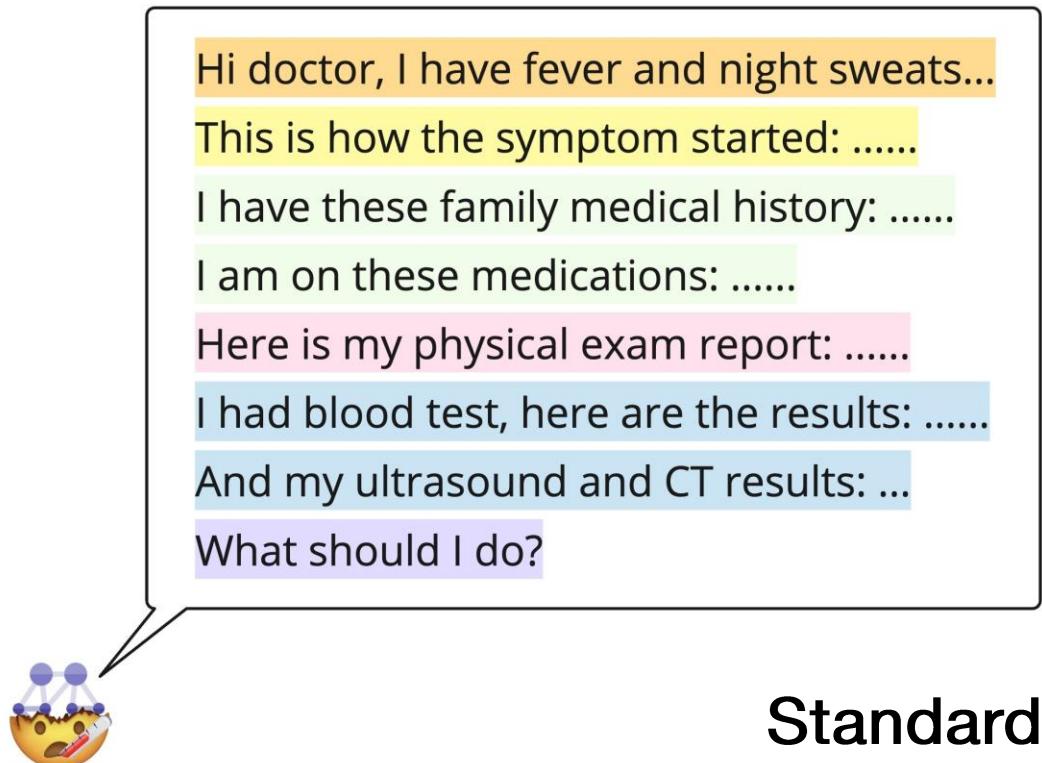


Stella Li

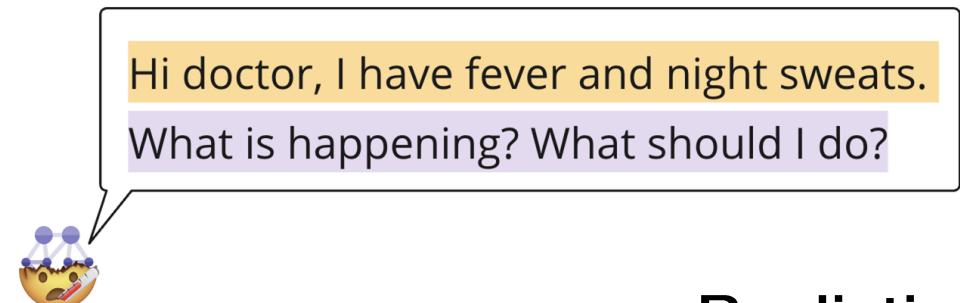
# Interaction data

- Most data in Web crawls are static
  - News
  - Books
  - Wikipedia
  - Papers
  - ...
- Even data from user interaction on sites like Reddit are not that interactive

# Our work: Evaluation in interactive settings

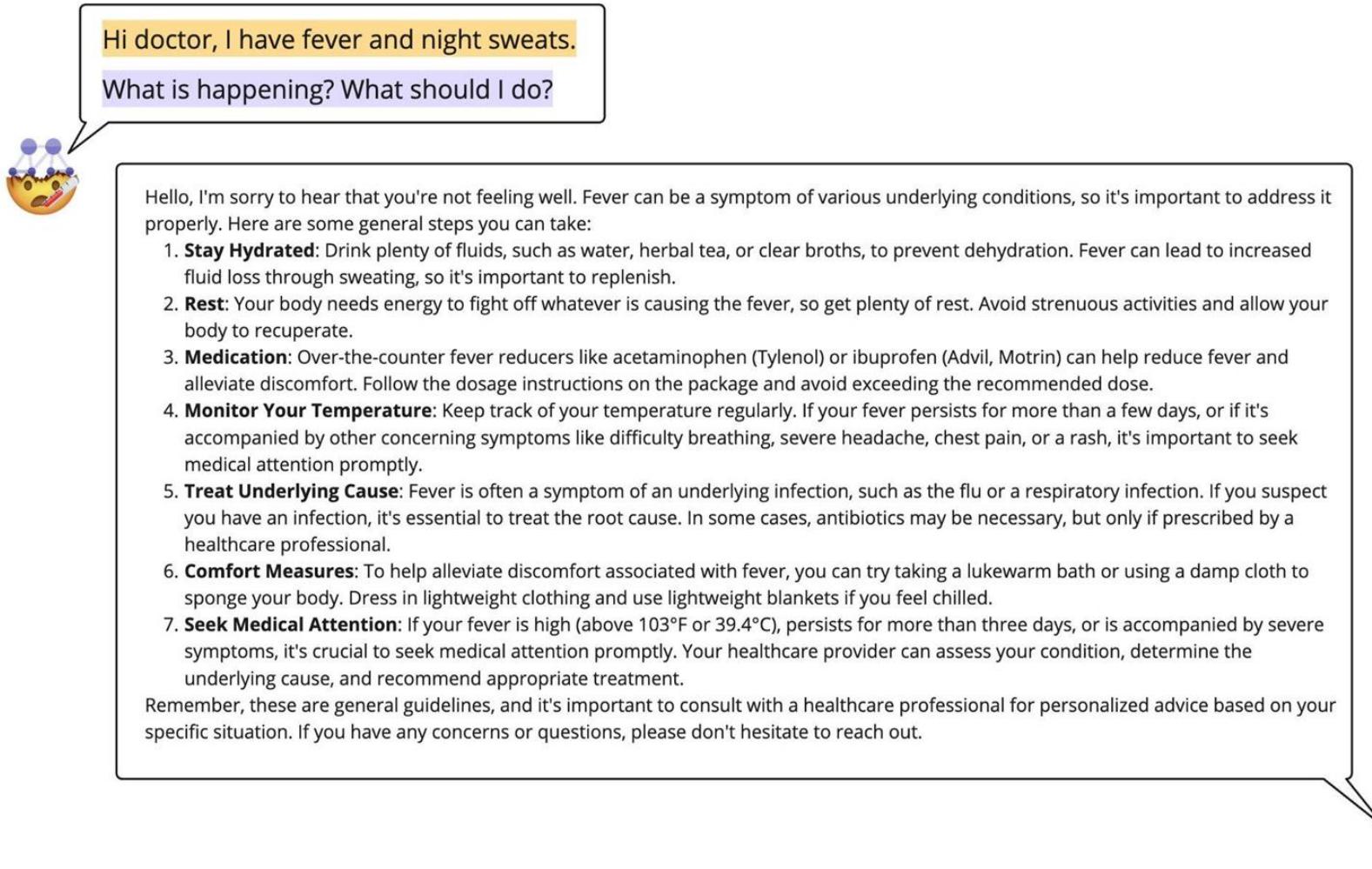


**Standard  
benchmarks**



**Realistic**

# How do state-of-the-art LLMs do?



Hi doctor, I have fever and night sweats.

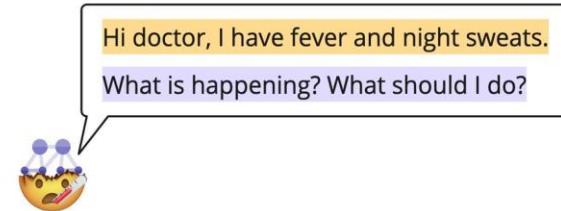
What is happening? What should I do?

Hello, I'm sorry to hear that you're not feeling well. Fever can be a symptom of various underlying conditions, so it's important to address it properly. Here are some general steps you can take:

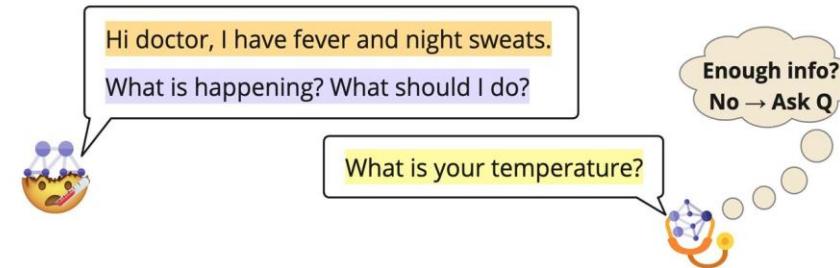
- Stay Hydrated:** Drink plenty of fluids, such as water, herbal tea, or clear broths, to prevent dehydration. Fever can lead to increased fluid loss through sweating, so it's important to replenish.
- Rest:** Your body needs energy to fight off whatever is causing the fever, so get plenty of rest. Avoid strenuous activities and allow your body to recuperate.
- Medication:** Over-the-counter fever reducers like acetaminophen (Tylenol) or ibuprofen (Advil, Motrin) can help reduce fever and alleviate discomfort. Follow the dosage instructions on the package and avoid exceeding the recommended dose.
- Monitor Your Temperature:** Keep track of your temperature regularly. If your fever persists for more than a few days, or if it's accompanied by other concerning symptoms like difficulty breathing, severe headache, chest pain, or a rash, it's important to seek medical attention promptly.
- Treat Underlying Cause:** Fever is often a symptom of an underlying infection, such as the flu or a respiratory infection. If you suspect you have an infection, it's essential to treat the root cause. In some cases, antibiotics may be necessary, but only if prescribed by a healthcare professional.
- Comfort Measures:** To help alleviate discomfort associated with fever, you can try taking a lukewarm bath or using a damp cloth to sponge your body. Dress in lightweight clothing and use lightweight blankets if you feel chilled.
- Seek Medical Attention:** If your fever is high (above 103°F or 39.4°C), persists for more than three days, or is accompanied by severe symptoms, it's crucial to seek medical attention promptly. Your healthcare provider can assess your condition, determine the underlying cause, and recommend appropriate treatment.

Remember, these are general guidelines, and it's important to consult with a healthcare professional for personalized advice based on your specific situation. If you have any concerns or questions, please don't hesitate to reach out.

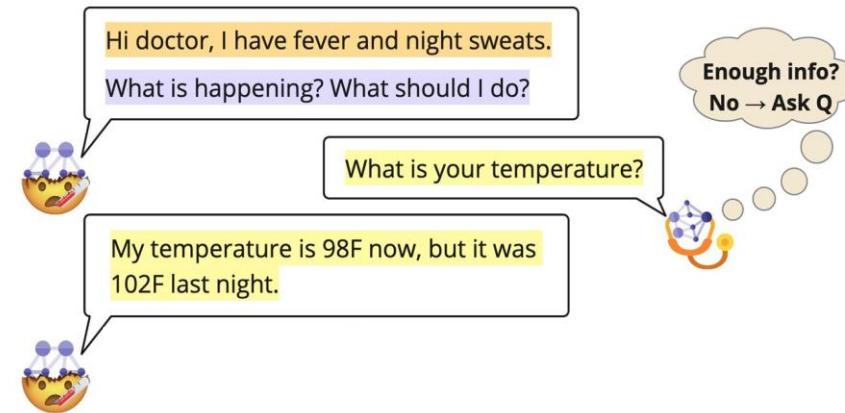
# A more realistic interaction



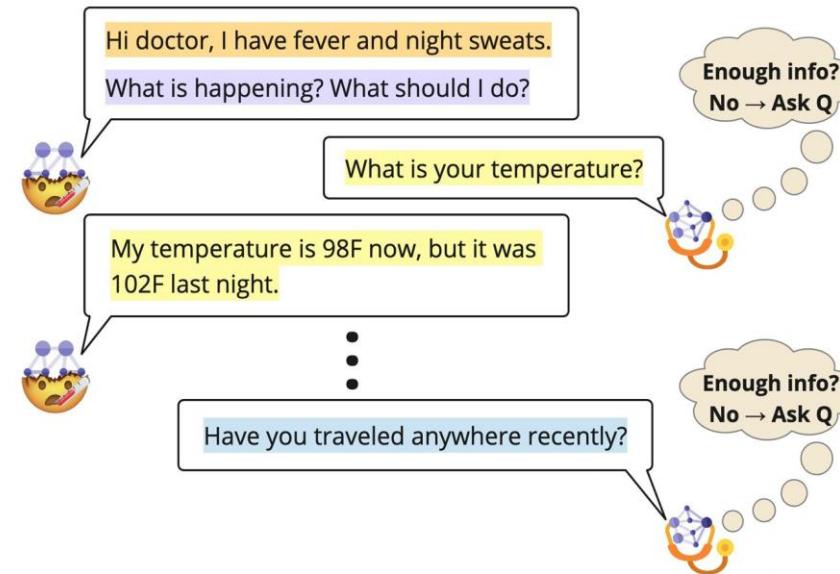
# A more realistic interaction



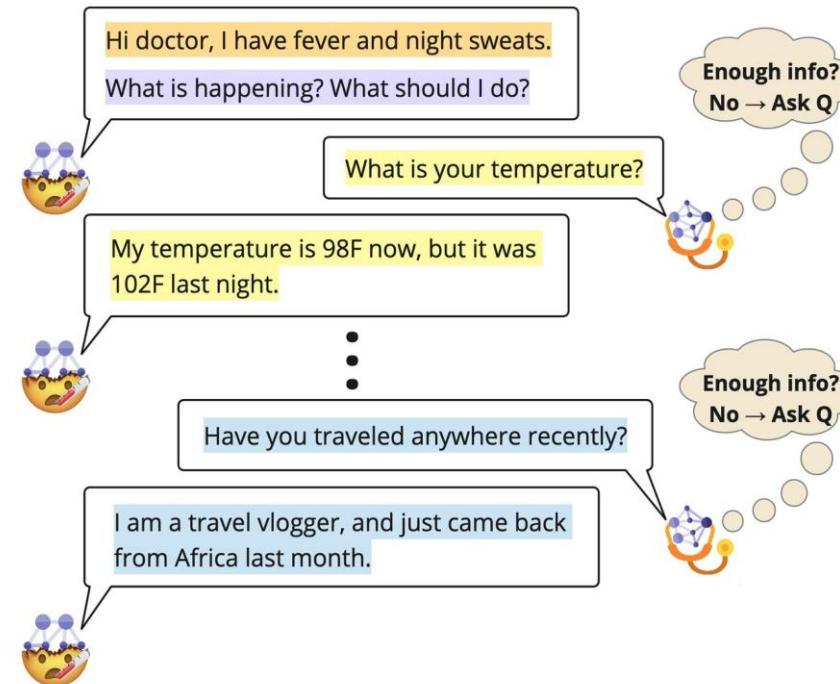
# A more realistic interaction



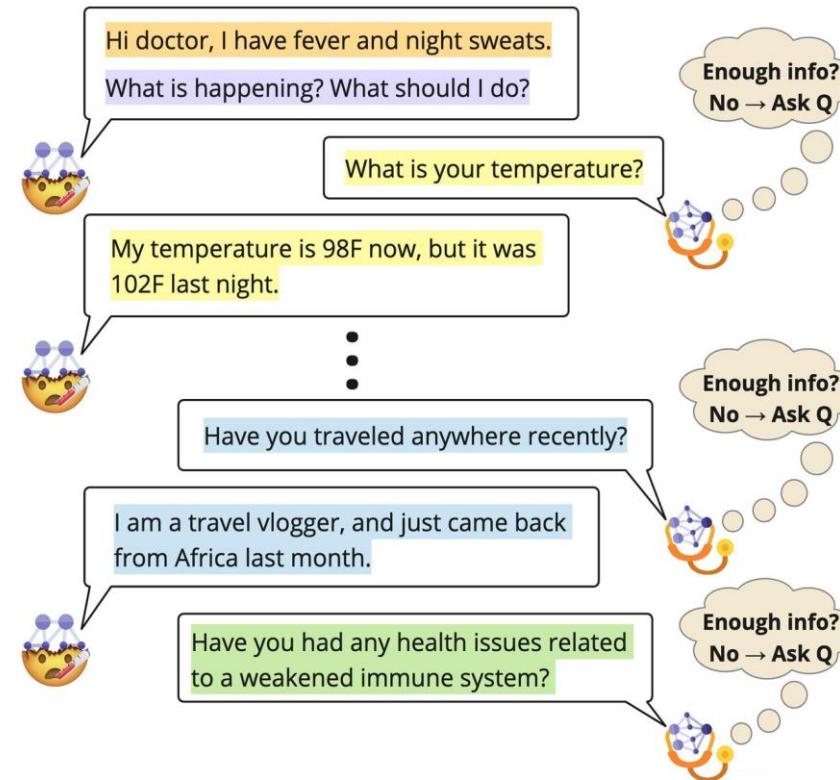
# A more realistic interaction



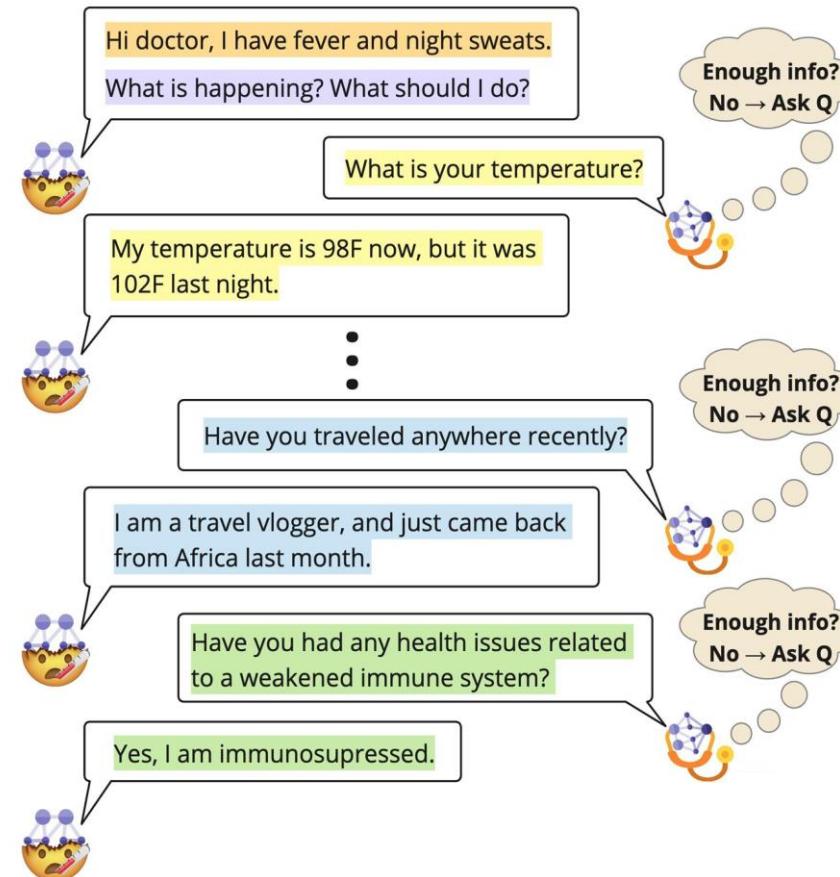
# A more realistic interaction



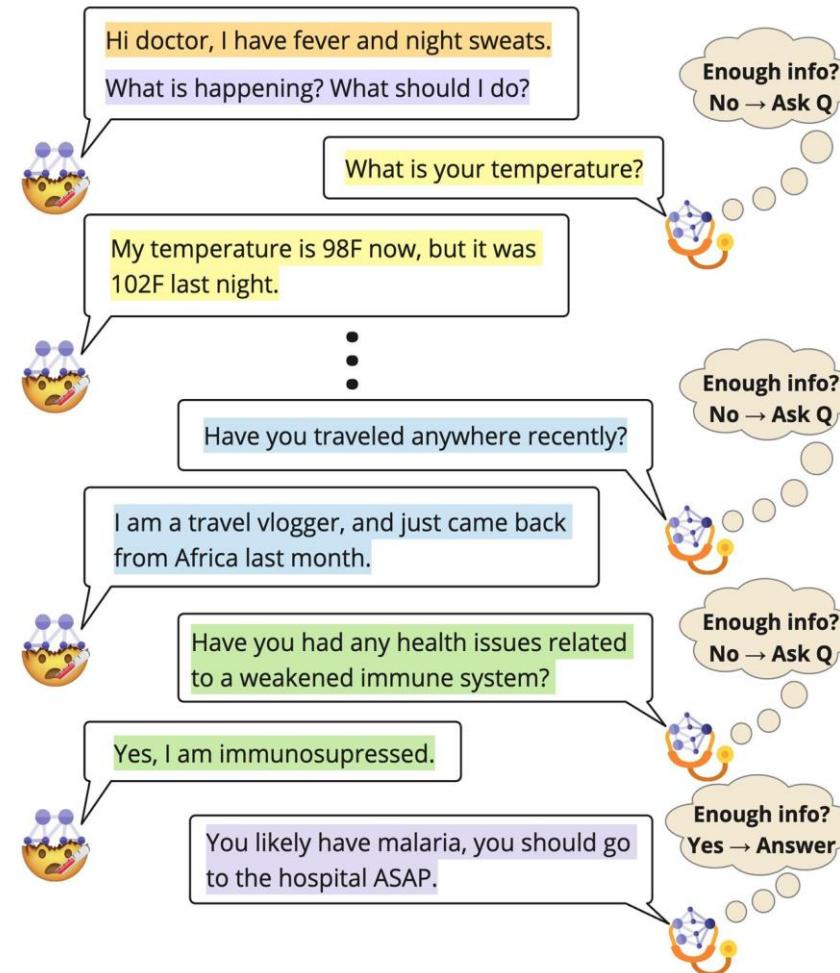
# A more realistic interaction



# A more realistic interaction



# A more realistic interaction



# Converting static to interactive benchmarks

- MedQA [1]: ~12k US Medical License Exam questions
- Craft-MD [2]: 140 skin disease case vignettes

**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

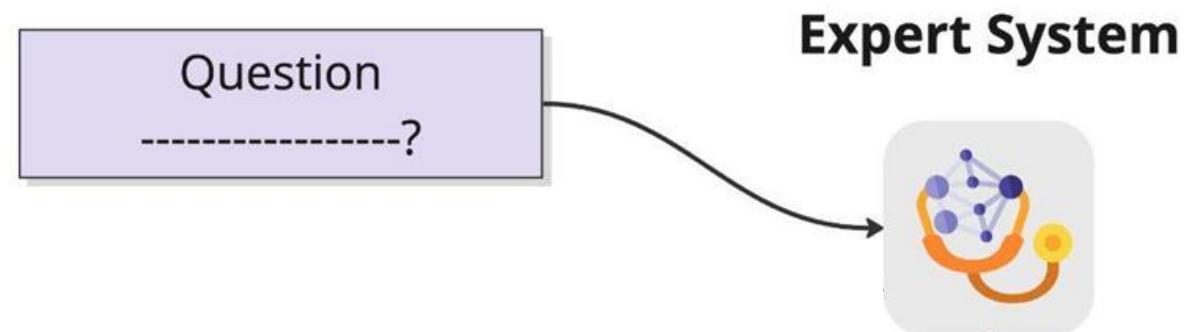
**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7C (98.6F), pulse 88/min. Laboratory results are significant for the following: | Hemoglobin 10.2 g/dL | Erythrocyte count 5.5 million/mm<sup>3</sup> | Mean corpuscular volume 65 µm<sup>3</sup> | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

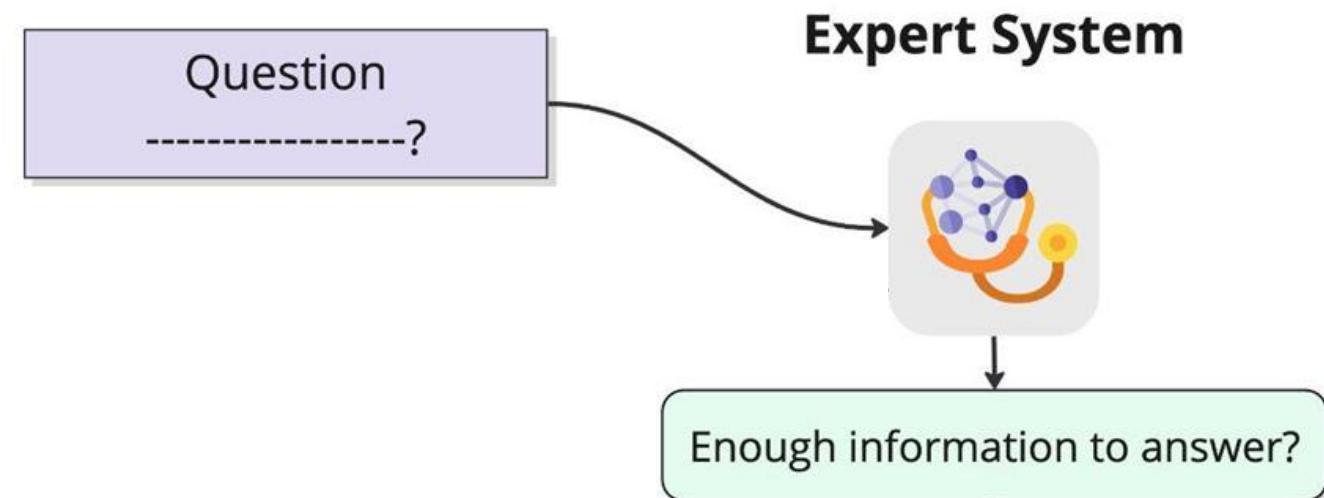
[1] Jin et al., 2020

[2] Johri et al., 2024

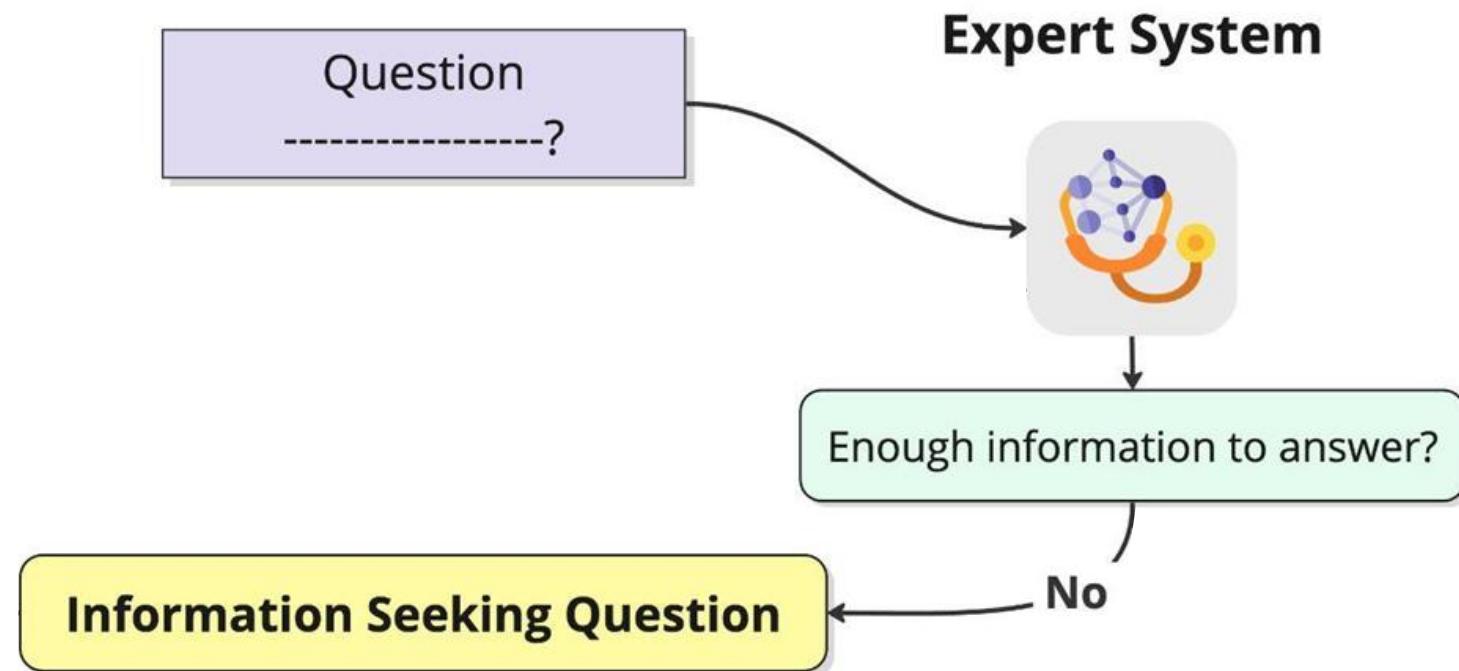
# The MEDIQ framework



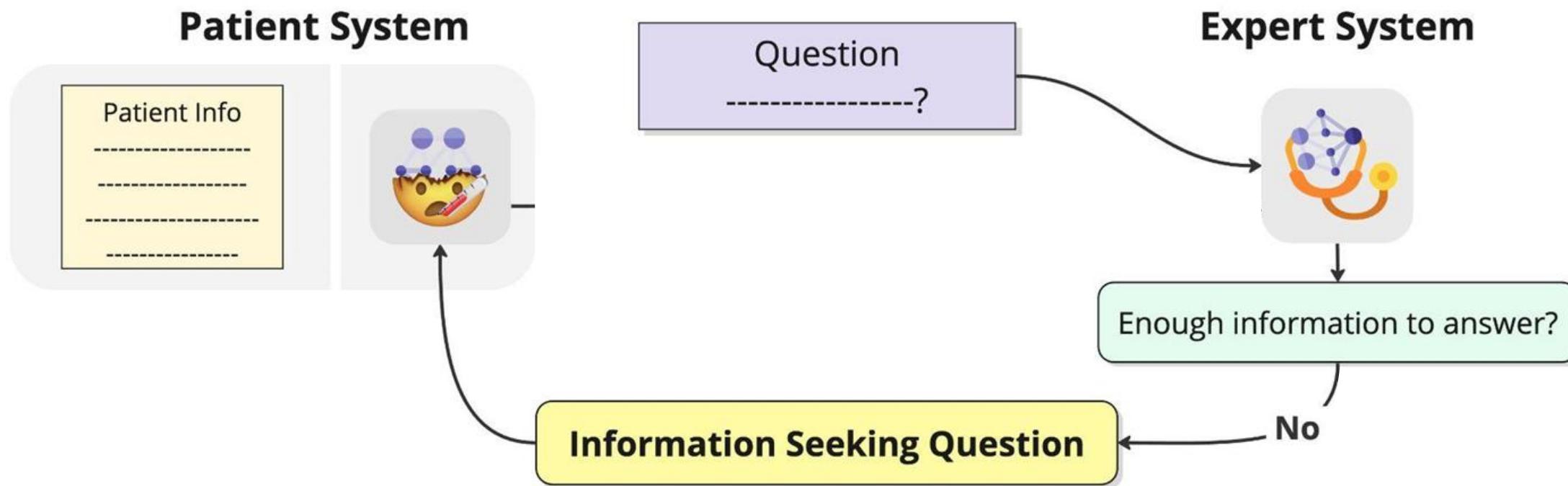
# The MEDIQ framework



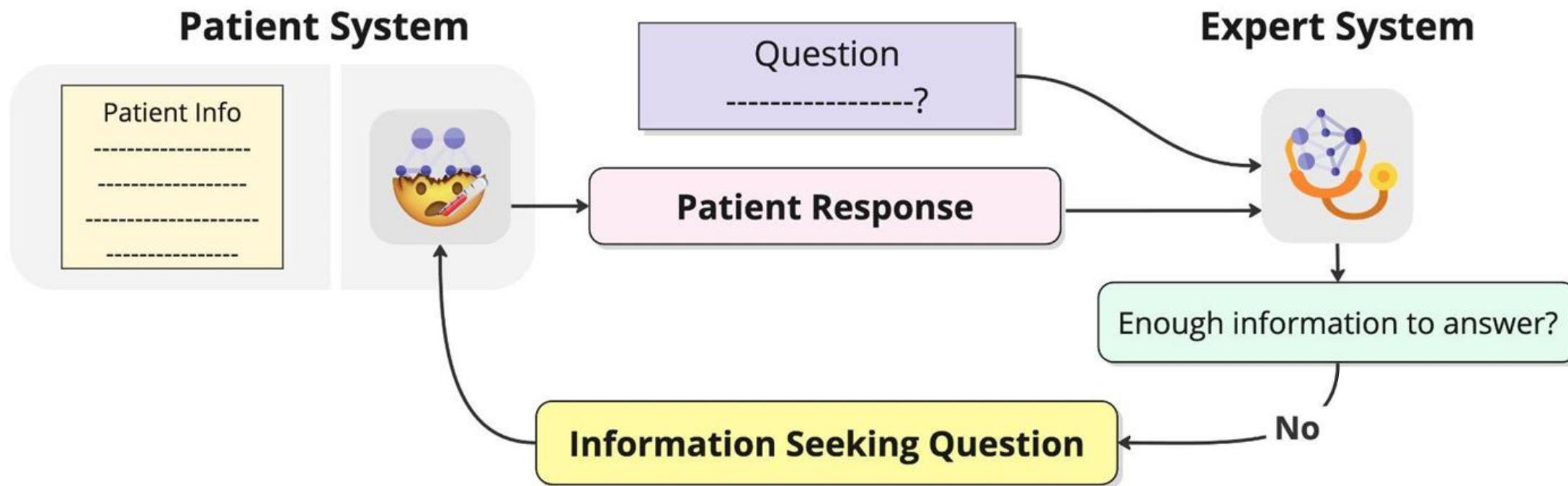
# The MEDIQ framework



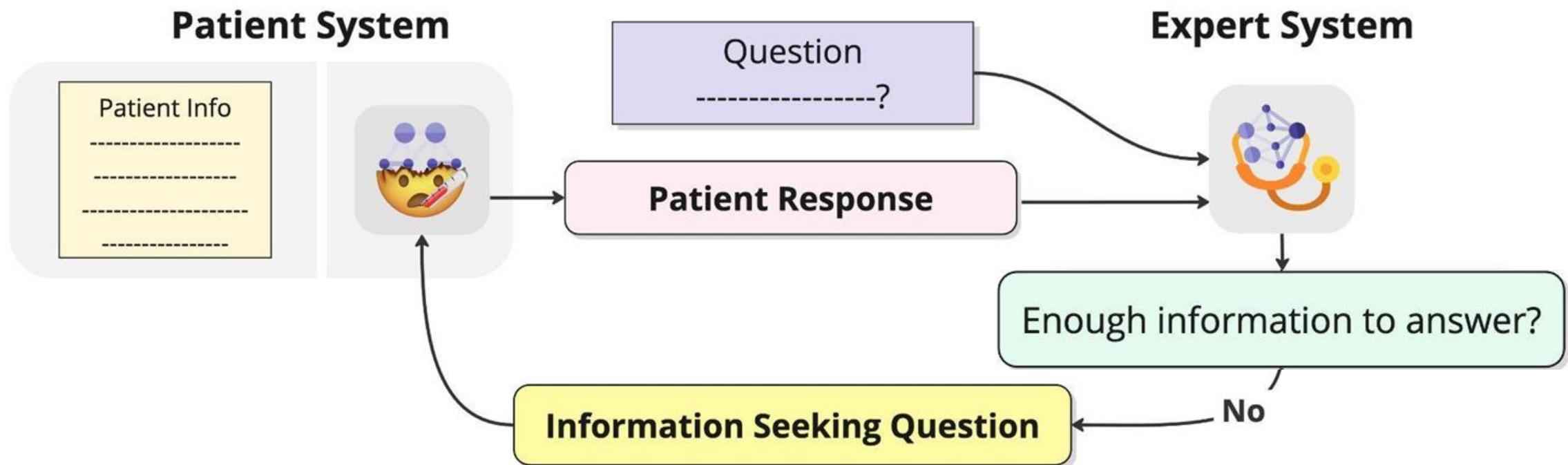
# The MEDIQ framework



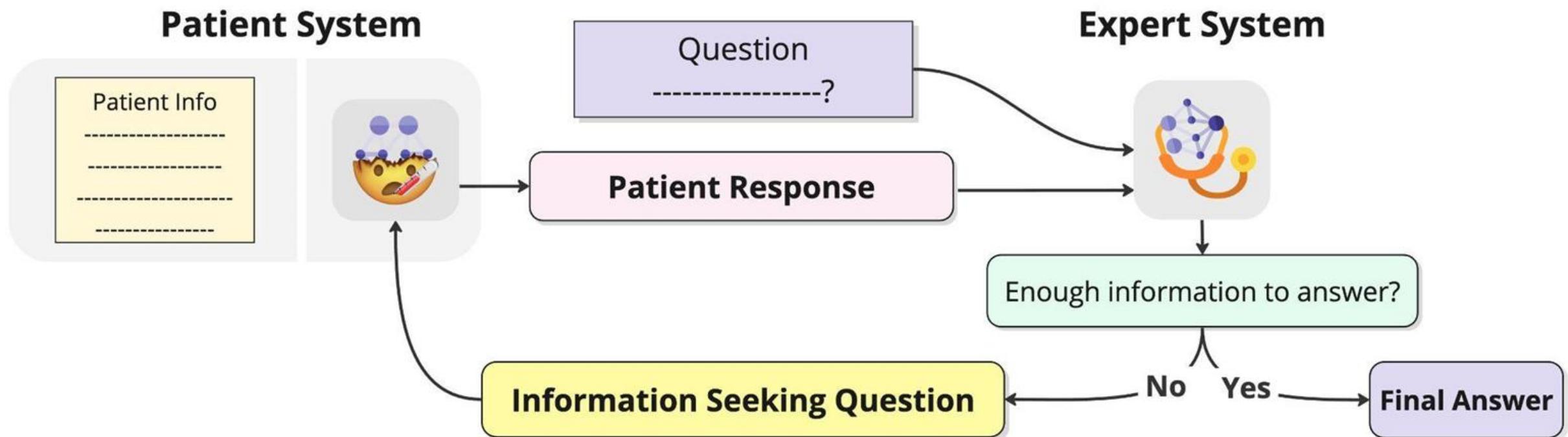
# The MEDIQ framework



# The MEDIQ framework



# The MEDIQ framework



<b>Task</b>	<b>Model</b>	<b>Non-Interactive Full</b>
<b>iMEDQA</b>	<b>Llama-2-70b</b>	42.88
	<b>Llama-3-8b</b>	68.06
	<b>Llama-3-70b</b>	84.66
	<b>GPT-3.5</b>	55.78
	<b>GPT-4</b>	79.70
<b>iCRAFT-MD</b>	<b>GPT-3.5</b>	82.14
	<b>GPT-4</b>	91.43

**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7C (98.6F), pulse 88/min. Laboratory results are significant for the following: | Hemoglobin 10.2 g/dL | Erythrocyte count 5.5 million/mm<sup>3</sup> | Mean corpuscular volume 65 µm<sup>3</sup> | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

Task	Model	Non-Interactive	
		Full	Limited
iMEDQA	Llama-2-70b	42.88	36.74
	Llama-3-8b	68.06	52.01
	Llama-3-70b	84.66	58.46
	GPT-3.5	55.78	45.55
	GPT-4	79.70	54.45
iCRAFT-MD	GPT-3.5	82.14	53.57
	GPT-4	91.43	67.86

**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7°C (98.6°F), pulse 88/min. Laboratory results are significant for the following: + Hemoglobin 10.2 g/dL + Erythrocyte count 5.5 million/mm³ + Mean corpuscular volume 65 µm³ + Mean corpuscular hemoglobin 21 pg/cell + Red cell distribution width 13.5% (ref: 11.5–14.5%)."

Task	Model	Non-Interactive			Interactive BASIC
		Full	Limited		
iMEDQA	Llama-2-70b	42.88	36.74	29.98	
	Llama-3-8b	68.06	52.01	33.04	
	Llama-3-70b	84.66	58.46	55.07	
	GPT-3.5	55.78	45.55	42.17	
	GPT-4	79.70	54.45	55.39	
iCRAFT-MD	GPT-3.5	82.14	53.57	45.00	
	GPT-4	91.43	67.86	73.57	



11% relative drop!

**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7C (98.6F), pulse 88/min. Laboratory results are significant for the following: | Hemoglobin 10.2 g/dL | Erythrocyte count 5.5 million/mm<sup>3</sup> | Mean corpuscular volume 65  $\mu$ m<sup>3</sup> | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

Task	Model	Non-Interactive		Interactive	
		Full	Limited	BASIC	BEST
iMEDQA	Llama-2-70b	42.88	36.74	29.98	<b>35.56</b>
	Llama-3-8b	68.06	52.01	33.04	<b>45.79</b>
	Llama-3-70b	84.66	58.46	55.07	<b>60.90</b>
	GPT-3.5	55.78	45.55	42.17	<b>50.20</b>
	GPT-4	79.70	54.45	55.39	<b>66.09</b>
iCRAFT-MD	GPT-3.5	82.14	53.57	45.00	<b>59.29</b>
	GPT-4	91.43	67.86	73.57	<b>84.29</b>

20% relative gap



**question:** "Which of the following is the most likely cause of this patient's anemia?"

**options:** (A) Vitamin B12 deficiency, (B) Gestational anemia, (C) Iron deficiency, (D) Thalassemia trait

**Initial Info:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit."

**context:** "A 27-year-old G1P0 woman at 9 weeks estimated gestational age presents for a prenatal visit. She is vegetarian and emigrated from Nepal 7 years ago. She does not use tobacco, alcohol or recreational drugs. The patient's vital signs include: blood pressure 111/95 mm Hg, temperature 36.7C (98.6F), pulse 88/min. Laboratory results are significant for the following: | Hemoglobin 10.2 g/dL | Erythrocyte count 5.5 million/mm<sup>3</sup> | Mean corpuscular volume 65  $\mu$ m<sup>3</sup> | Mean corpuscular hemoglobin 21 pg/cell | Red cell distribution width 13.5% (ref: 11.5-14.5%)."

# Conformal inference

- Instead of predicting a single label, predict a set of potential labels
- Guaranteed coverage: % that true label is in the prediction set

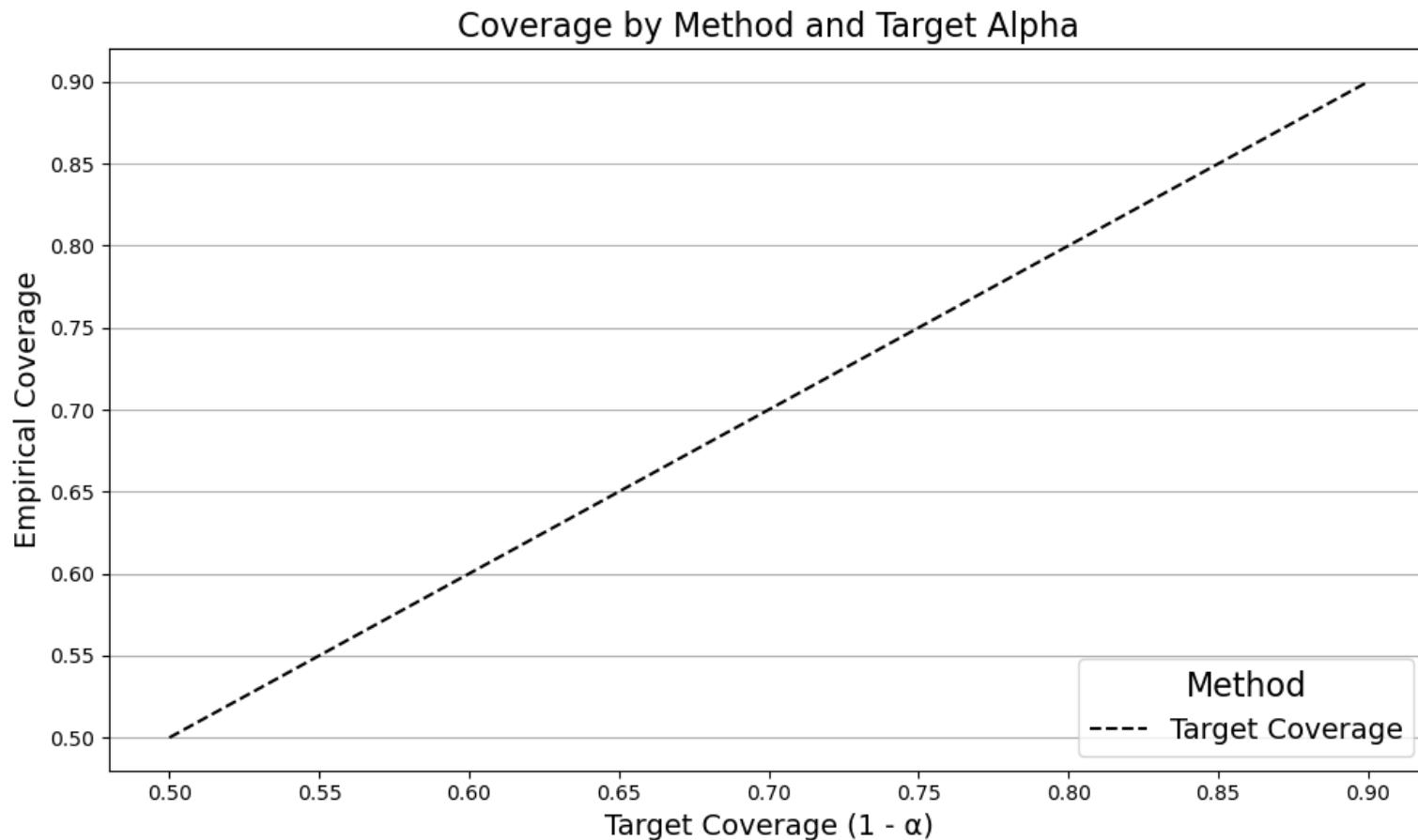


Eric Frankel

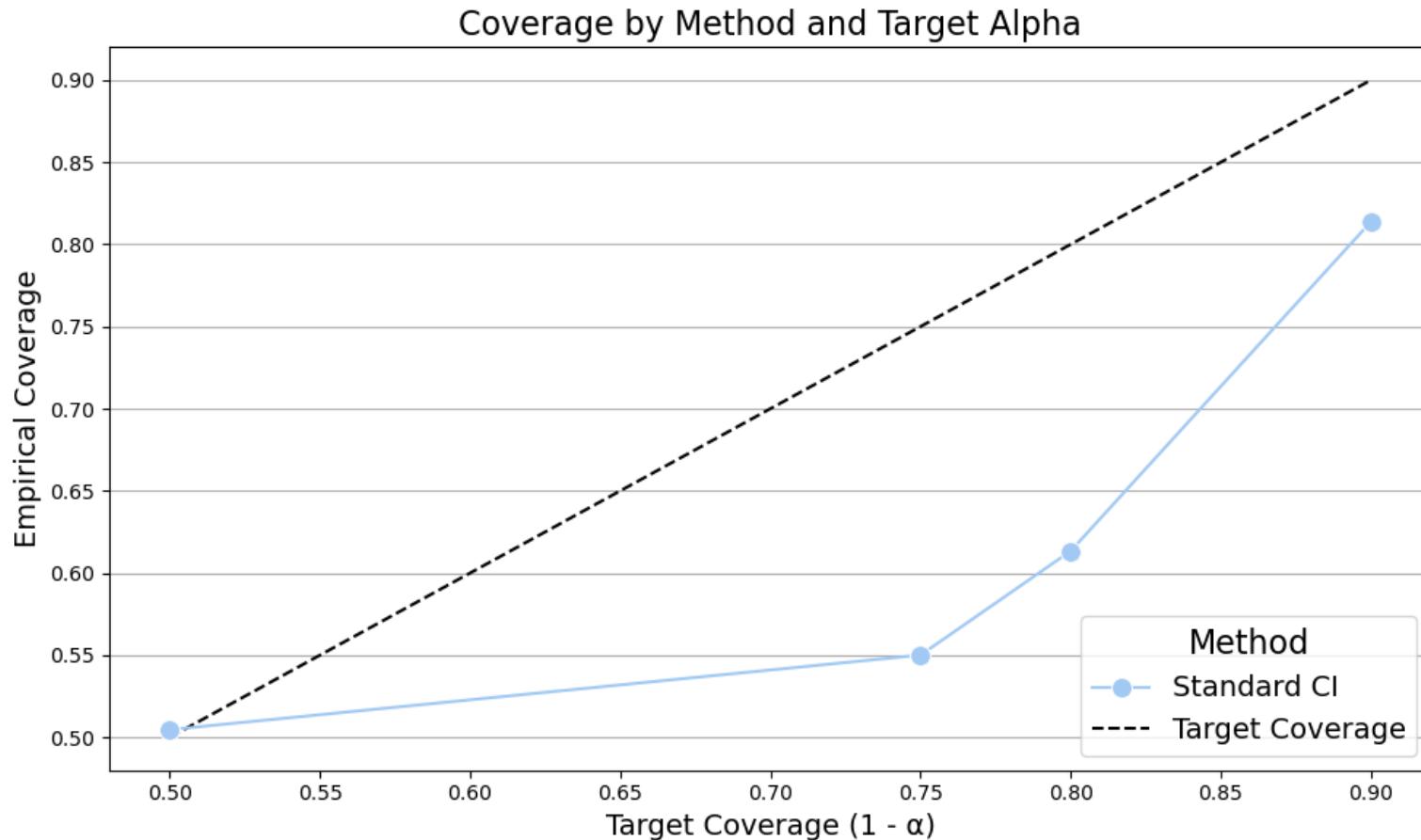


Stella Li

# Conformal inference



# Conformal inference



# Takeaways

- Many tasks are naturally interactive
- Whereas current data/benchmarks are largely static
- More work needed to fill this gap
  - Amie from Google [McDuff et al., 2024]
  - Uncertainty of Thoughts [Hu et al., 2024]

Try MEDIQ out: <https://stellalisy.com/projects/mediQ/>

# This talk: Reliable & responsible data use

1. Scaling up data at inference time
2. Training on synthetic data
3. Evaluating on interaction data

pangwei@cs.washington.edu



Rulin Shao



Jacqueline He



Scott Geng



Stella Li