

Foundation Models and Prompts

Danna Gurari

University of Colorado Boulder
Spring 2025



<https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/>

Review

- Last lecture:
 - Other data modalities
 - Internet-scale trained models
 - Scaling laws
 - Lab assignment 3: multimodal task
 - Programming tutorial
- Assignments (Canvas):
 - Lab assignment 3 due in 1 week
- Questions?

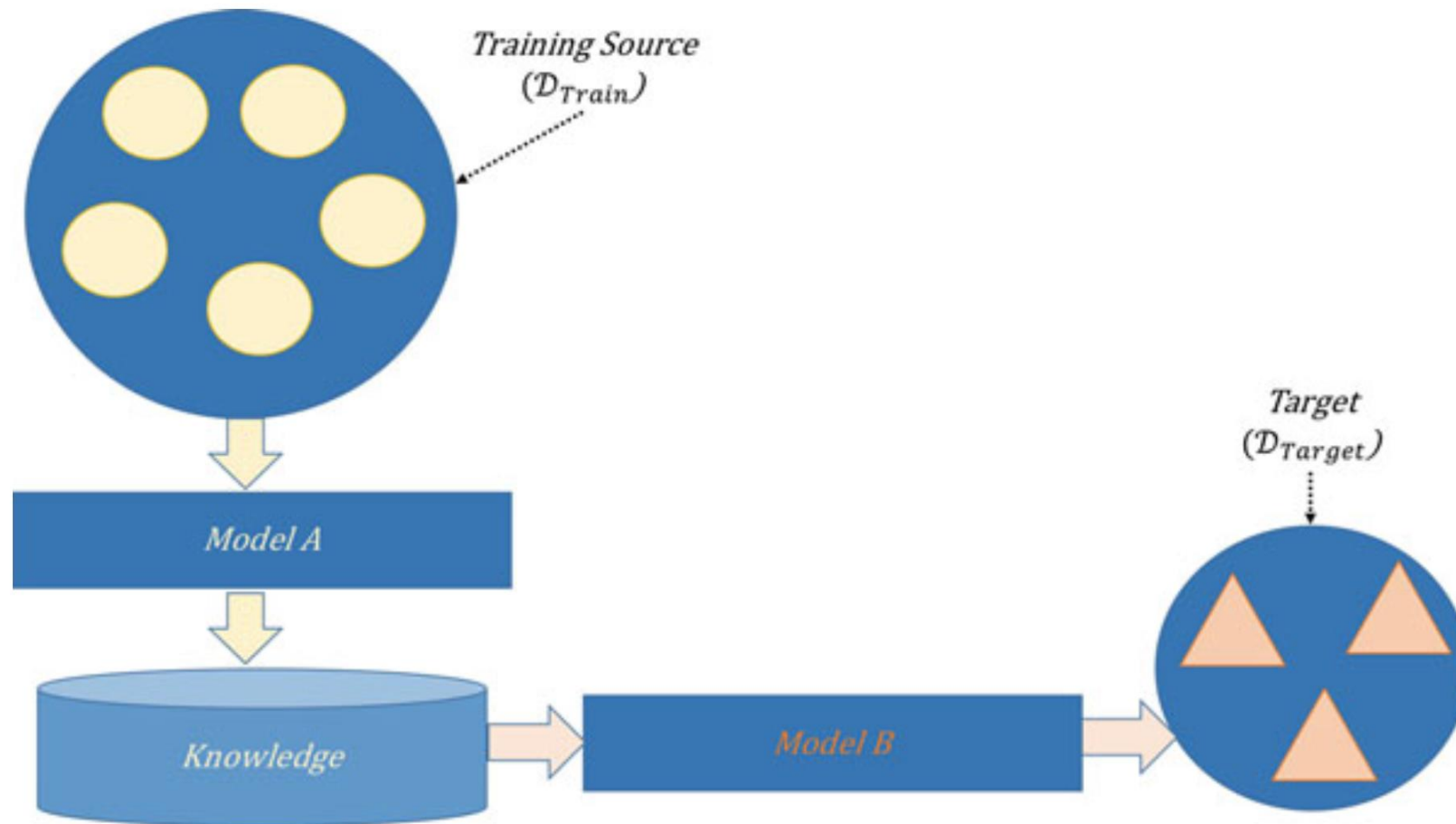
Today's Topics

- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

Today's Topics

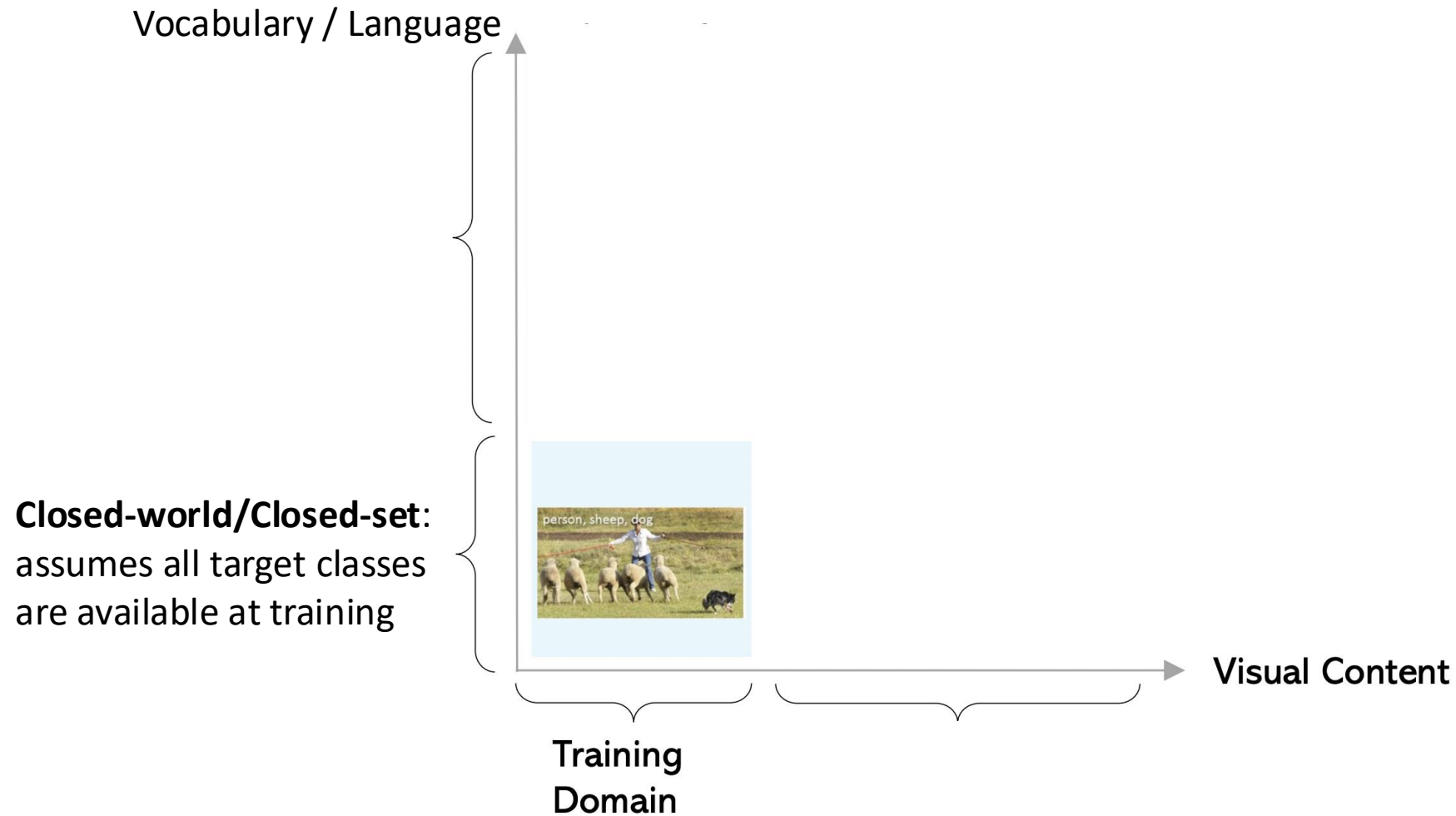
- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

Focus for 2010s: Closed-World Problems



Learning works well with “big” labeled datasets for the target task!

Open Problems: Beyond Closed-World Setting

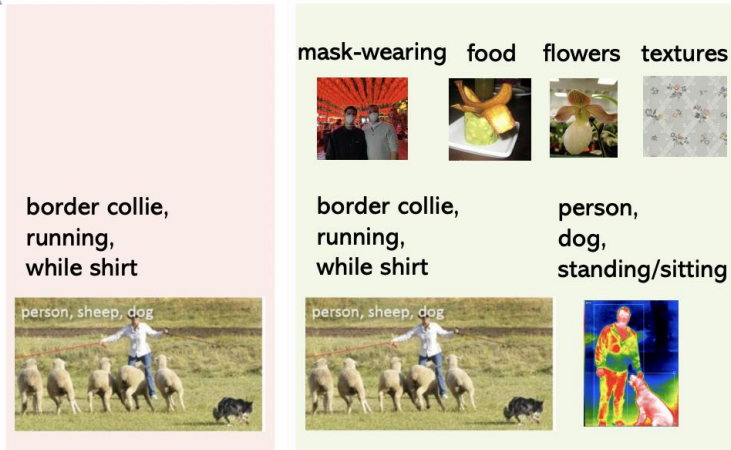


Open Problems: Beyond Closed-World Setting

Vocabulary / Language

Open vocabulary and Zero-shot:
generalize to task with no labeled training data for the target task (e.g., novel categories), where the former problem permits annotations with novel category (for a different task)

Closed-world/Closed-set:
assumes all target classes are available at training



Open world/In the wild for different tasks (e.g., detection):
succeed for all categories, whether seen or not seen during training

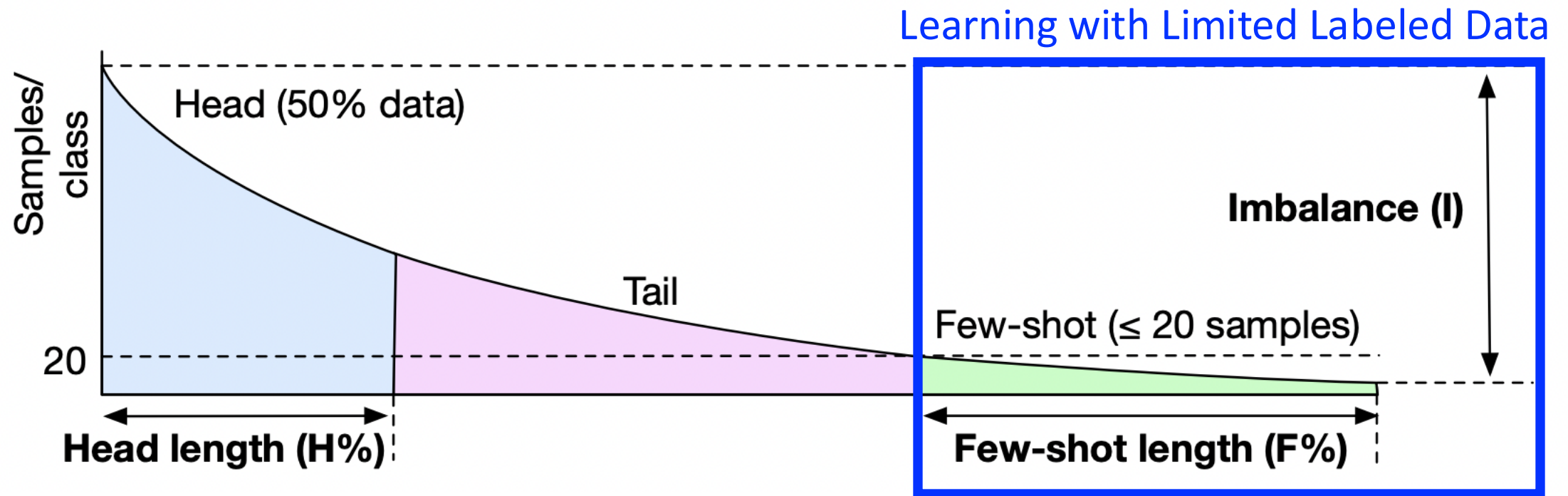
Training Domain

Out-of-domain/Robustness Testing:
same content observed differently

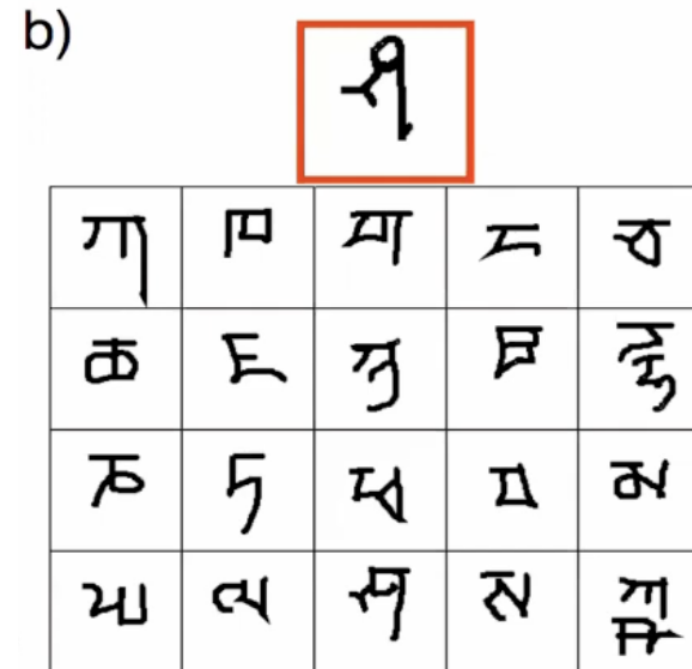
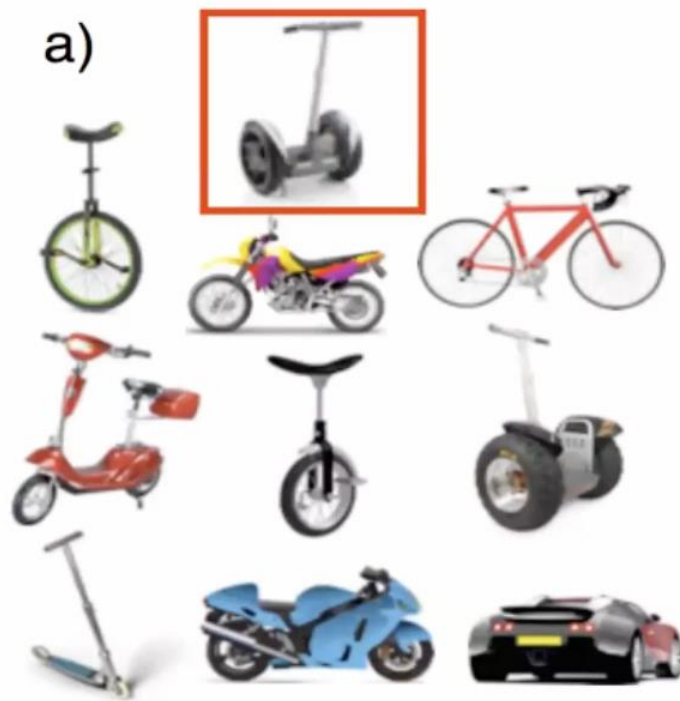
Visual Content

Open set classification/Out-of-distribution Detection:
predict whether a sample is drawn from the distribution observed at training time

Beyond “Big” Data: Few-Shot Learning



Beyond “Big” Data: Few-Shot Learning Intuition



Lake et al, 2013, 2015

Given one example per category, identify the category of the **query**

Beyond “Big” Data: Zero-Shot Learning Intuition



What is this?

How many examples do you think you would need to see to recognize one of these?

Beyond “Big” Data: Zero-Shot Learning Intuition



Could see 0 examples if you knew the object fuses a person on top with a horse on the bottom

Beyond “Big” Data: Zero-Shot Learning Intuition



Could see 0 examples of a zebra if you knew it looks like a horse with black and white stripes

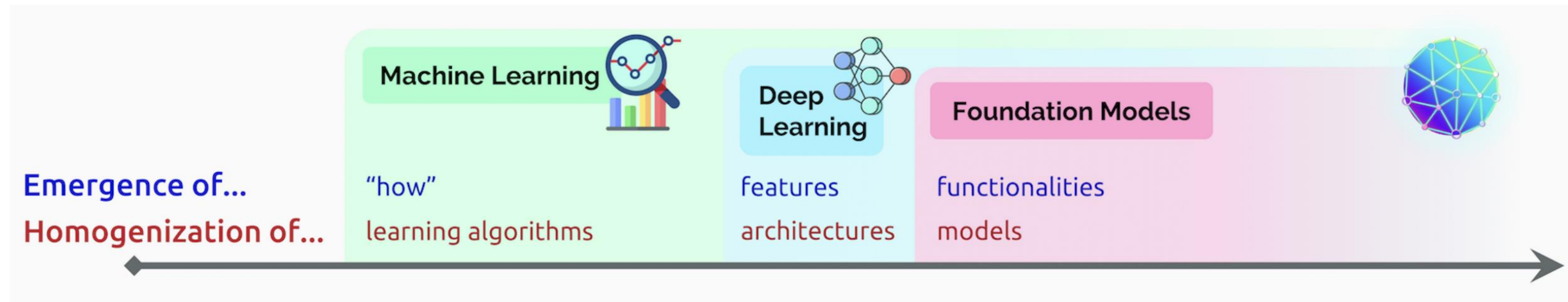
Today's Topics

- Motivation
- **Foundation Models**
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

New Paradigm:

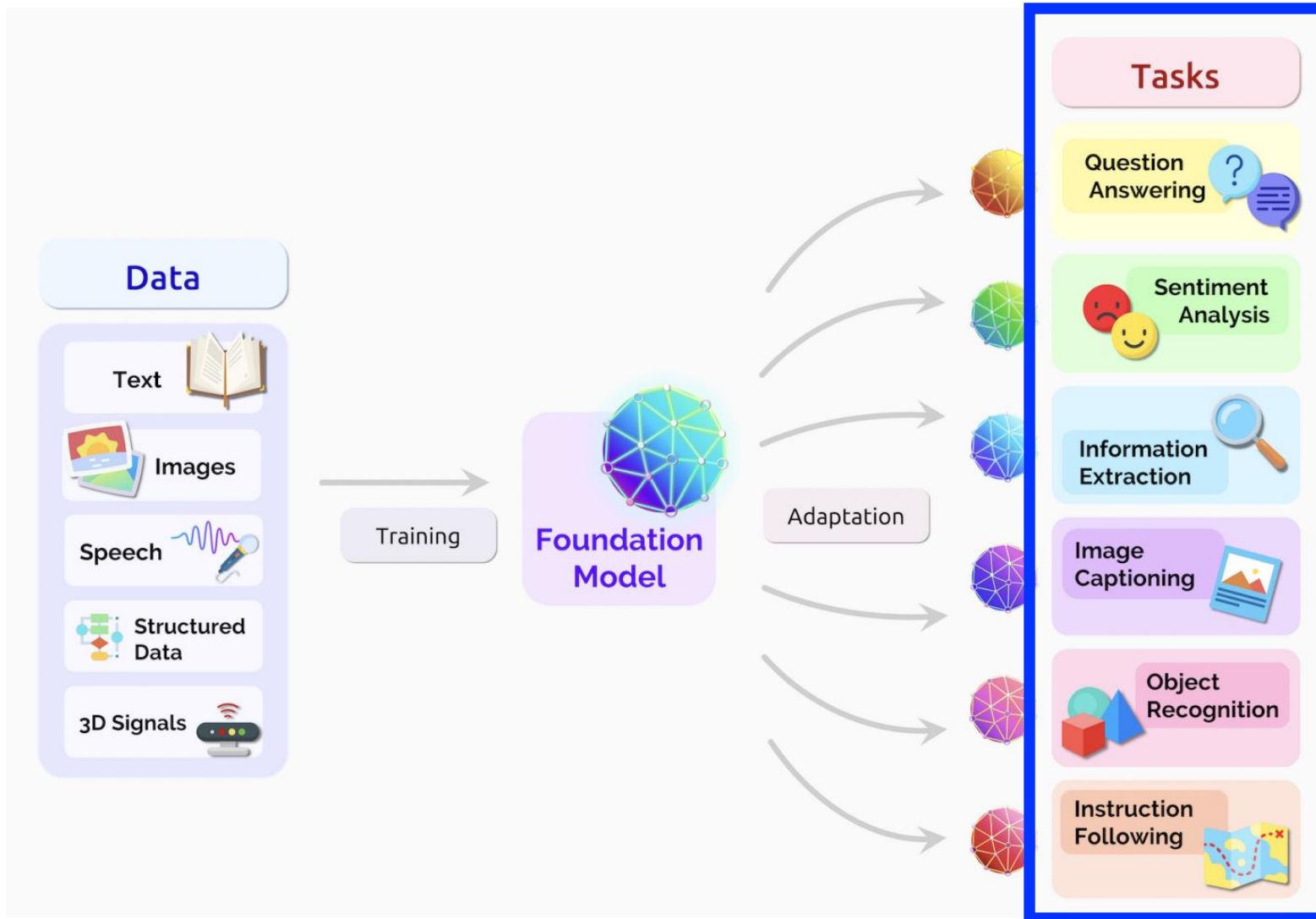
Foundation Models Can Generalize Beyond
Closed-World Settings With Limited Training Data

Definition of “Foundation Model”



Coined in 2021, it references the recent paradigm shift to develop a single model that can implicitly support many downstream tasks.

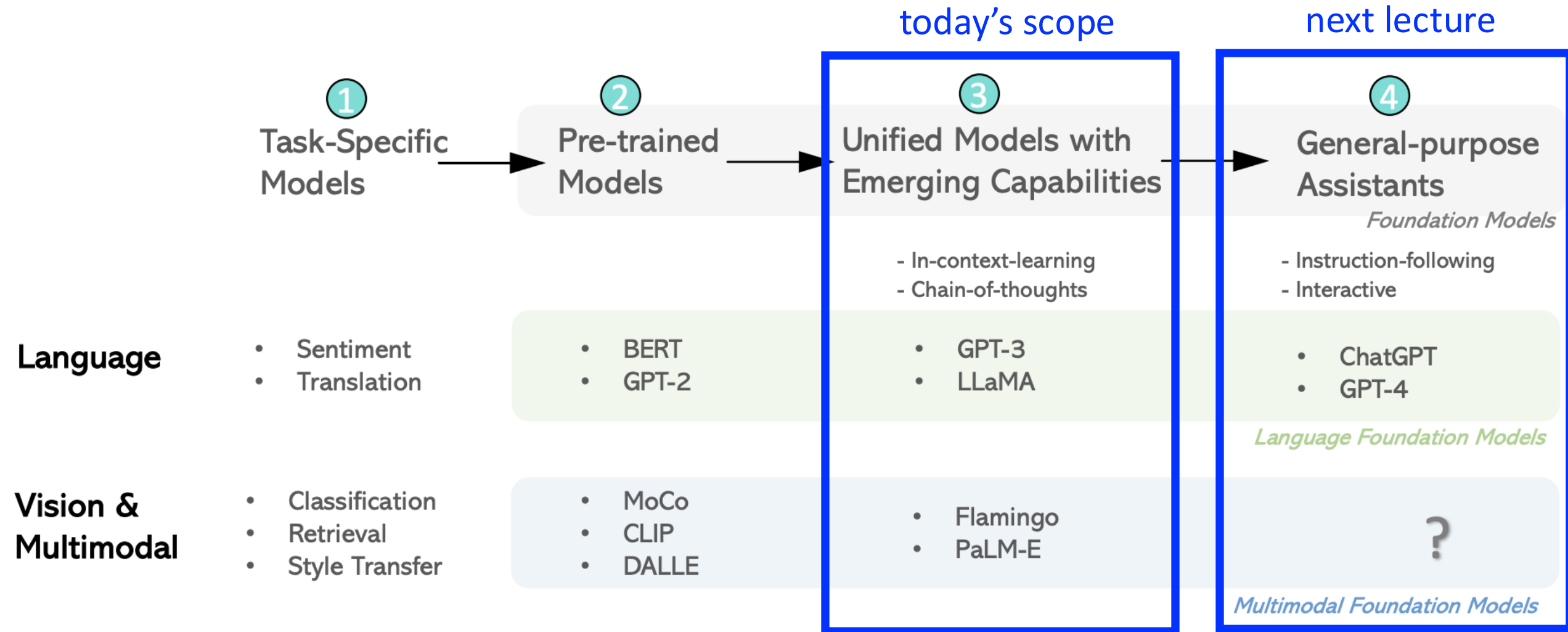
Training and Evaluating Foundation Models



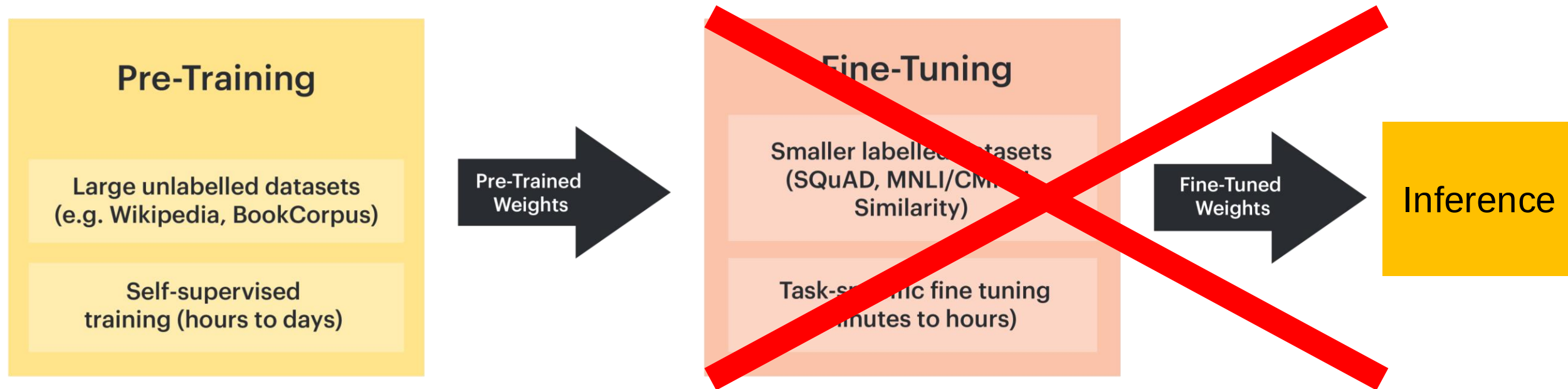
Evaluate with modern benchmark datasets for many:

1. **Different tasks** (e.g., object recognition, scene classification)
2. **Different distributions of the same task** (e.g., ImageNet vs data from blind people)

A Transition from Specialist to General-Purpose Models

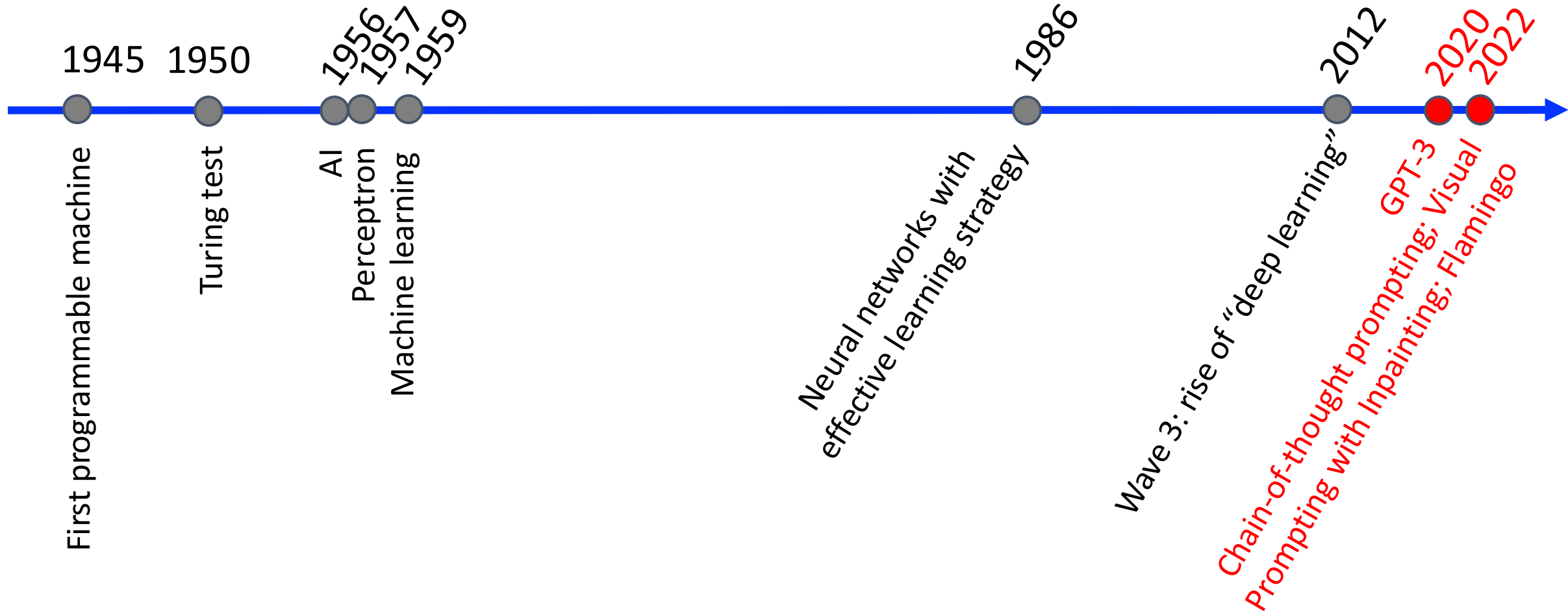


Emergent Abilities from Foundation Models



In 2020, we observed a model can be used *as is* for many downstream tasks with just *prompting*!

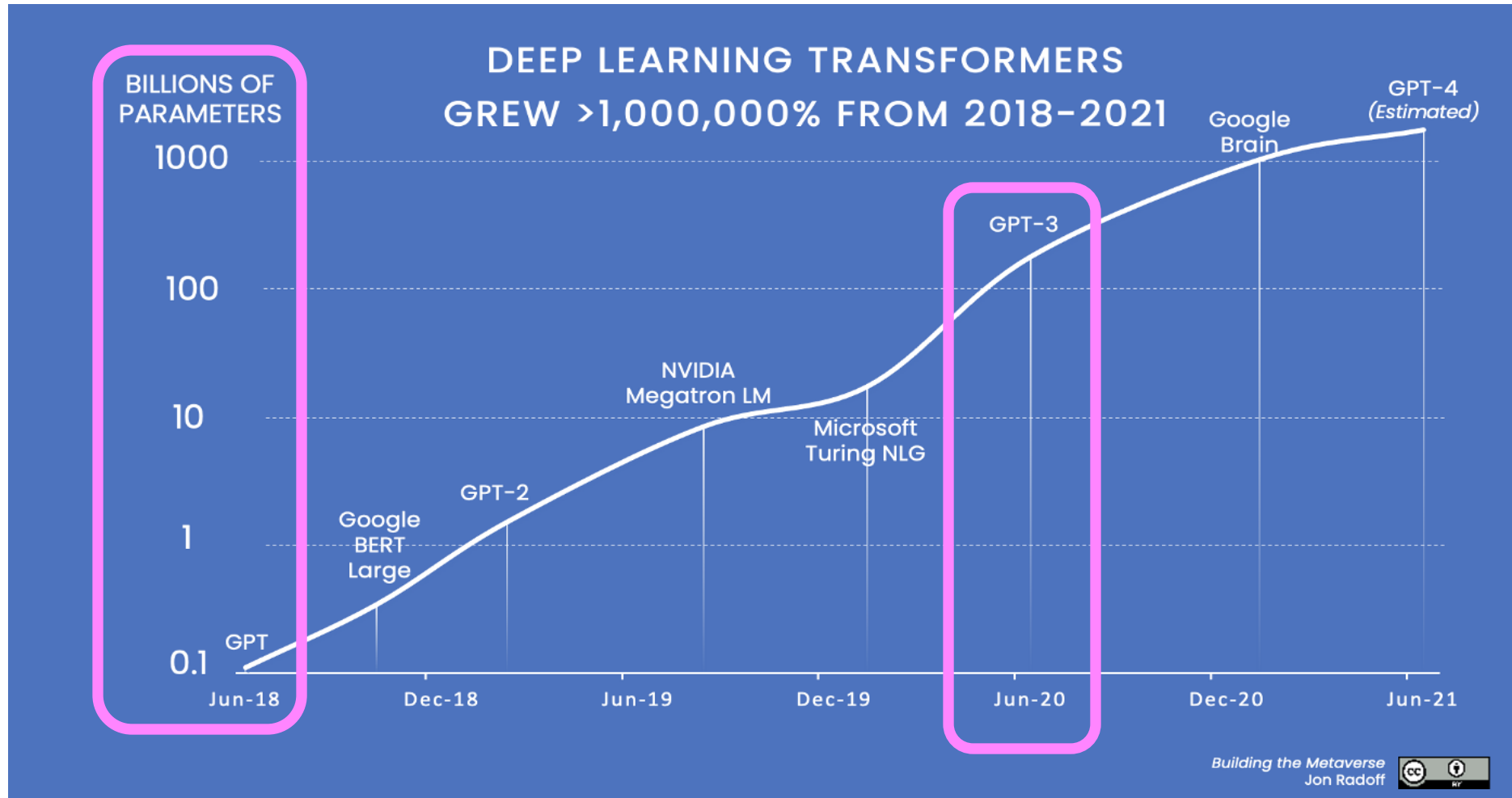
Pioneering Methods: Historical Context



Today's Topics

- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

Starting Point: Better Performance from Increasing Model and Dataset Sizes (Scaling Law)



(GPT-3; Neurips 2020)

Language Models are Few-Shot Learners

Tom B. Brown*

Benjamin Mann*

Nick Ryder*

Melanie Subbiah*

Jared Kaplan[†]

Prafulla Dhariwal

Arvind Neelakantan

Pranav Shyam

Girish Sastry

Amanda Askell

Sandhini Agarwal

Ariel Herbert-Voss

Gretchen Krueger

Tom Henighan

Rewon Child

Aditya Ramesh

Daniel M. Ziegler

Jeffrey Wu

Clemens Winter

Christopher Hesse

Mark Chen

Eric Sigler

Mateusz Litwin

Scott Gray

Benjamin Chess

Jack Clark

Christopher Berner

Sam McCandlish

Alec Radford

Ilya Sutskever

Dario Amodei

OpenAI

Architecture: Scaled-Up GPT-2 (aka GPT-3)

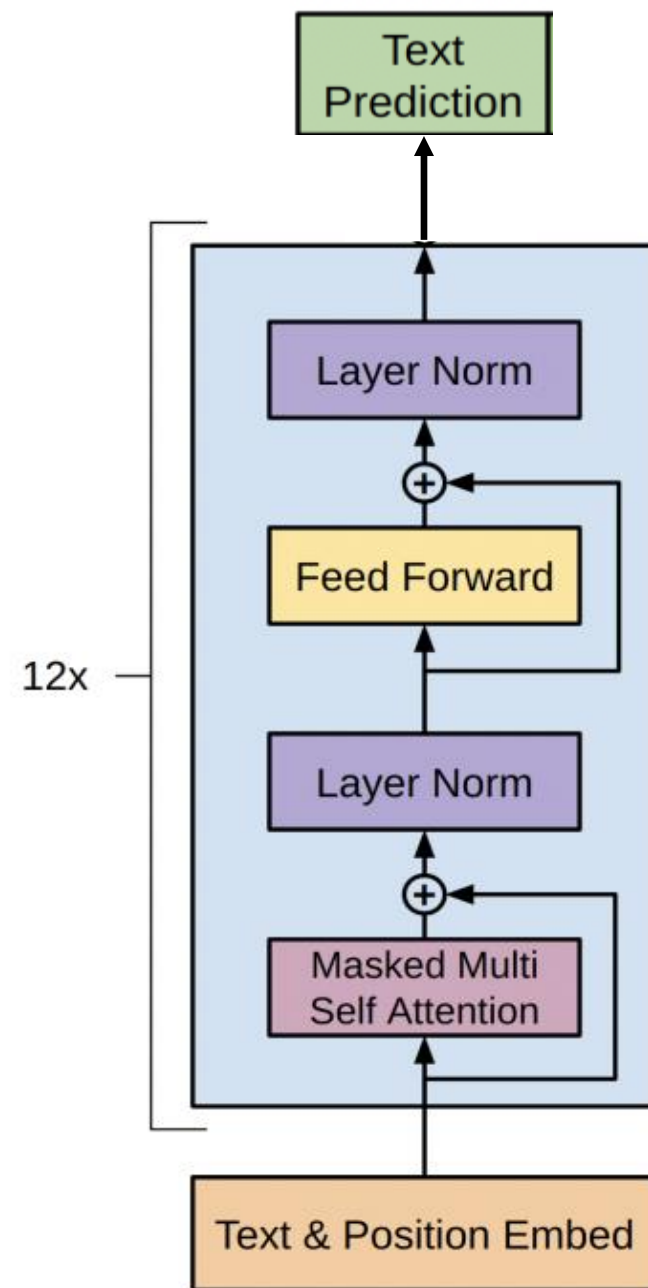
GPT-2 architecture with minor modifications; e.g.,

- More layers (96 vs 48) and so parameters (175B vs 1.5B)
- Accepts up to **2,048 input tokens** instead of 1,024 tokens.

Is the input size equivalent to approximately:

- (a) 1/3 single-space page
- (b) 1 single-space page
- (c) 3 single-space pages

Is this model (a) **discriminative** or (b) **generative**?



Training Data: Scaled-Up from GPT-2

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

(**Noisy, lower quality** Common Crawl data spanning 41 monthly crawls from 2016 to 2019, supplemented with 4 **high-quality datasets**)

Language composition: by word count, 93% English

How many GB for training? ~ (a) 6, (b) 60, (c) 600, (d) 6,000

Could the entire dataset fit on your personal computer?

- Not mine (it has ~250GB)

(web archives without HTML markup and non-text content <https://commoncrawl.org/>)

A screenshot of the Common Crawl website. The header features the 'COMMON CRAWL' logo on the left and navigation links 'The Data' and 'Resources' on the right. The main content area includes a paragraph stating 'Common Crawl is a 501(c)(3) non-profit founded in 2007. We make wholesale extraction, transformation and analysis of open web data accessible to researchers.' Below this is an 'Overview' button. Further down, four key statistics are listed: 'Over 250 billion pages spanning 17 years.', 'Free and open corpus since 2007.', 'Cited in over 10,000 research papers.', and '3-5 billion new pages added each month.'

COMMON CRAWL

The Data ▾ Resources ▾

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

Overview

Over **250 billion** pages spanning 17 years.

Free and open corpus since 2007.

Cited in over **10,000** research papers.

3-5 billion new pages added each month.

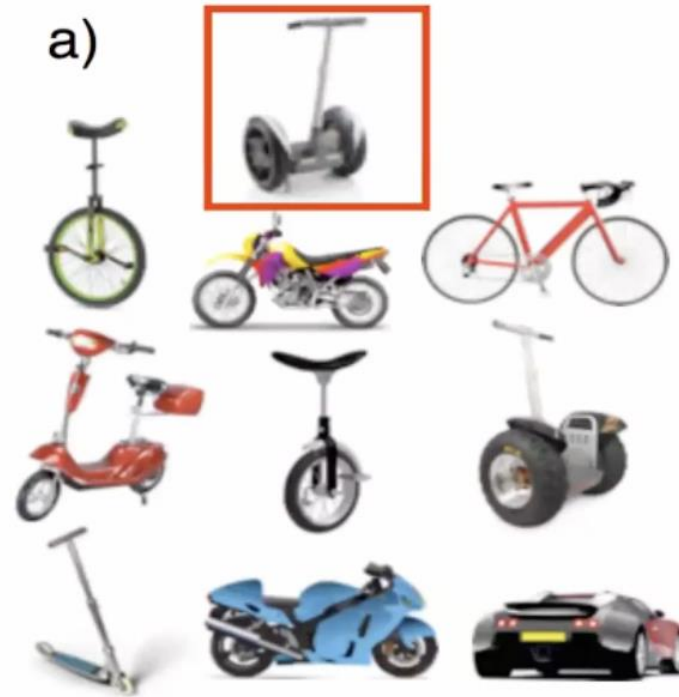
Training Protocol

- Same objective as GPT (**predict next word**)
- **Higher quality datasets sampled more** often (and so potential overfitting)
- Trained on V100 GPUs for $3.14\text{E}+23$ flops (floating point operations)
- What do you think is the most accurate estimate for the **training cost**?
 - (a) ~\$5,00
 - (b) ~\$50,000
 - (c) ~\$500,000
 - (d) ~\$5,000,000
- Required ~1.3 Gigawatt-hours of electricity (enough to power 121 homes in America for a year)

Two Key Emergent Abilities

- **In-context learning**: model completes document in format modeled by task-specific examples
- **Chain-of-thought (CoT) prompting**: model conveys its reasoning process when completing the task

Intuition: Can **You** Learn to Recognize Something With Zero/Few Examples?



<https://www.youtube.com/watch?v=9j4iH9TPTd8>

So Can GPT-3 with Prompts: Instructions + Examples (Latter Called “In-Context Learning”)

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

(Typically uses maximum amount of examples that can fit in 2,048 sized context window: ~10 to 100)

Prompt Designed Per Dataset; e.g.,

Context →	Q: What school did burne hogarth establish?
	A:
Target Completion →	School of Visual Arts

Figure G.35: Formatted dataset example for WebQA

Context →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. =
Target Completion →	In no case may they be used for commercial purposes.

Figure G.36: Formatted dataset example for De→En. This is the format for one- and few-shot learning, for this and other language tasks, the format for zero-shot learning is “Q: What is the {language} translation of {sentence} A: {translation}.”

Example Result: Fake News Generation

Title: United Methodists Agree to Historic Split

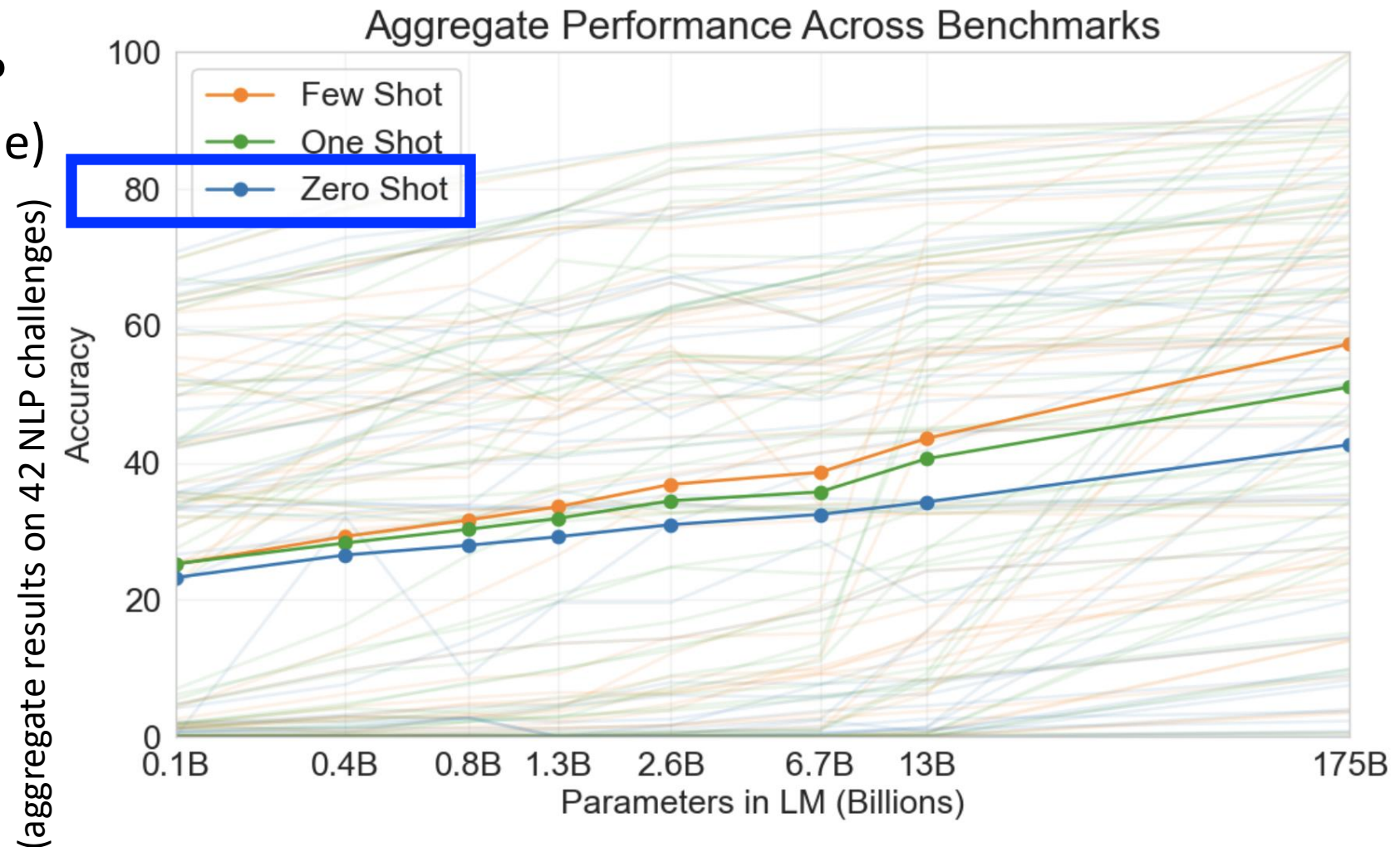
Subtitle: Those who oppose gay marriage will form their own denomination

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

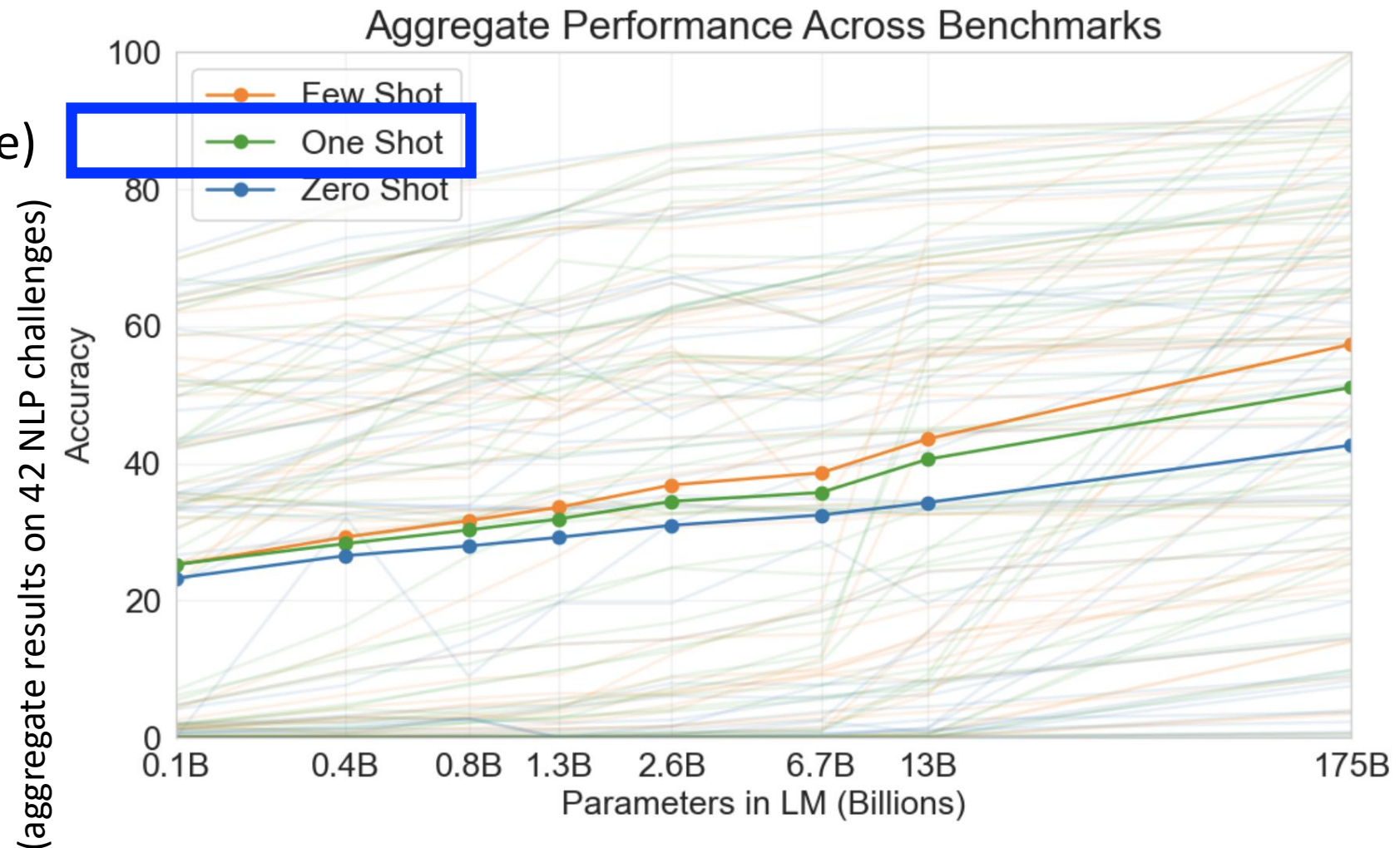
Results: Smooth Scaling with Model Capacity

What is the trend for zero-shot performance?
(recall zero-shot example)



Results: Smooth Scaling with Model Capacity

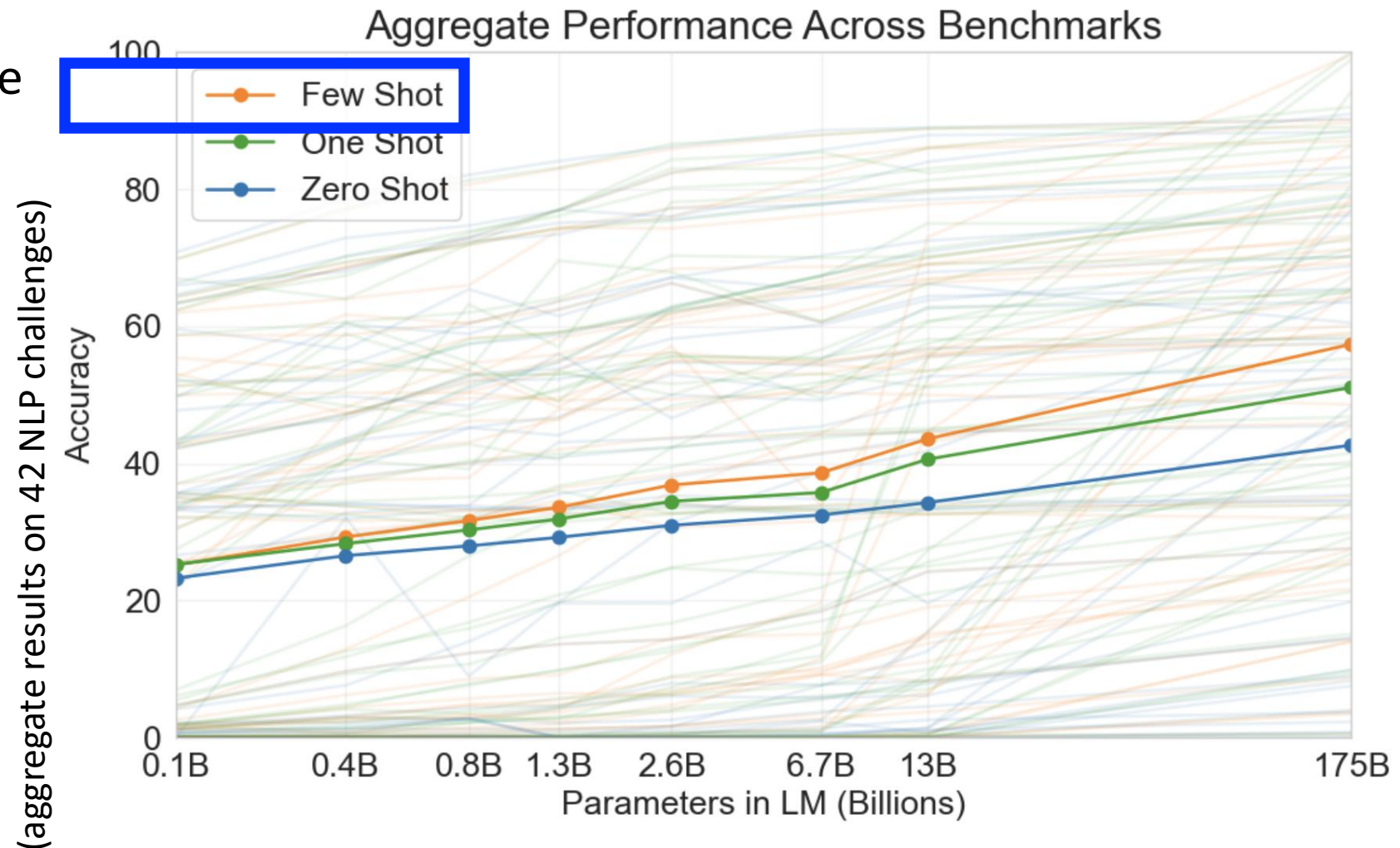
What is the trend for one-shot performance? (recall one-shot example)



Results: Smooth Scaling with Model Capacity

What are trends from providing more than one example to the model?

Performance gap between 3 settings grows, suggesting larger models learn **BETTER** from examples



Summary of GPT-3 Findings

Tested on **10s of NLP datasets**, showing strong performance overall and occasionally **state-of-the-art performance** (outperforming fine-tuned methods)!

Two Key Emergent Abilities

- **In-context learning**: model completes document in format modeled by task-specific examples
- **Chain-of-thought (CoT) prompting**: model conveys its reasoning process when completing the task

Intuition: How Much Reasoning Do You Do?

- What is $2 + 2$?
 - 4
- What is 24×14 ?
 - My reasoning process:
 - $24 \times 10 = 240$
 - $24 \times (14-10) = 24 \times 4 = 96$
 - $240+96 = 336$ (final answer!)
 - Did you follow the same or a different reasoning path to generate the answer?
- We can also encourage a model to reason/think before responding

(Pioneering paper; Neurips 2022)

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

CoT Prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Guides model to show intermediate reasoning steps

CoT Prompting: Why Does it Work?

CoT

Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.

Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✓

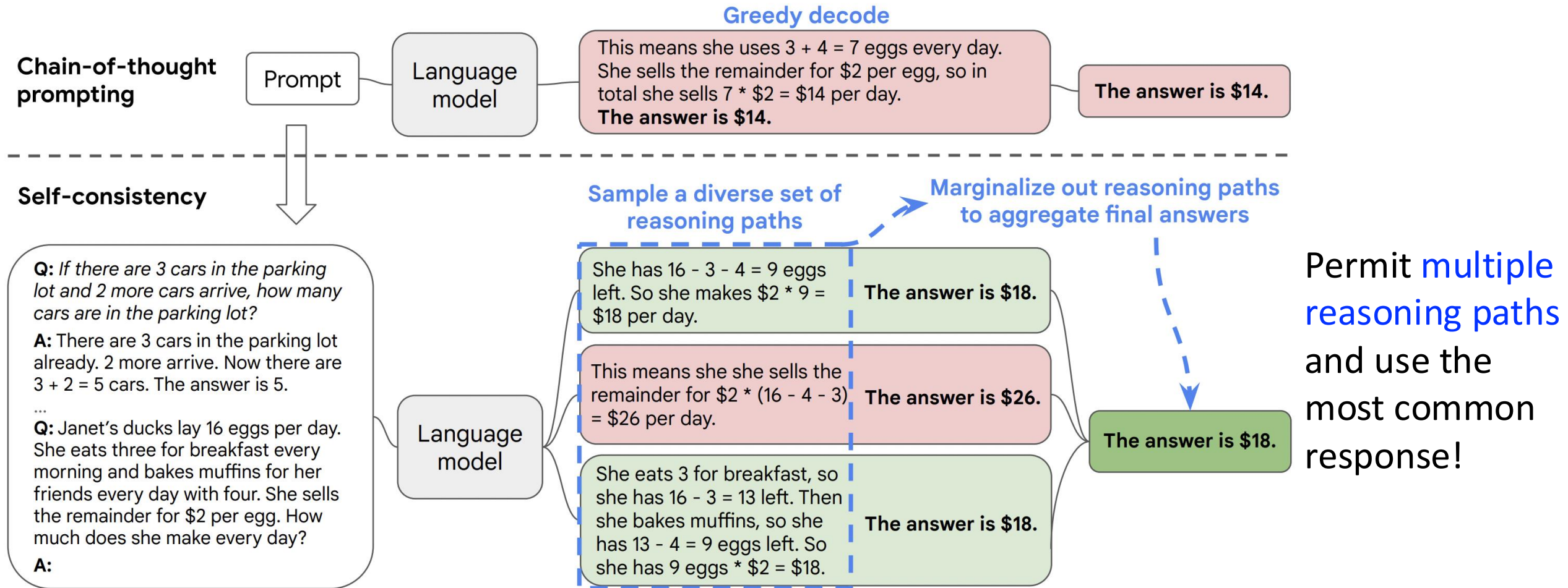
Performance can still improve with **invalid examples**, suggesting benefit of examples is revealing target format

Invalid Reasoning

Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39.

Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✓

CoT Prompting Enhancement: Self-Consistency



CoT Prompting Enhancement: Self-Consistency

Introduces diversity by **randomly sampling**, at each time step, the next token **from top K most probable tokens** (from softmax layer)

Sample a diverse set of reasoning paths

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.

...

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?

A:

Language model

She has $16 - 3 - 4 = 9$ eggs left. So she makes $\$2 * 9 = \18 per day.

The answer is \$18.

This means she she sells the remainder for $\$2 * (16 - 4 - 3) = \26 per day.

The answer is \$26.

She eats 3 for breakfast, so she has $16 - 3 = 13$ left. Then she bakes muffins, so she has $13 - 4 = 9$ eggs left. So she has $9 \text{ eggs} * \$2 = \18 .

The answer is \$18.

CoT Variant

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. **X**

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. **✓**

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. **✓**

Models perform better even when just asked to reason slowly

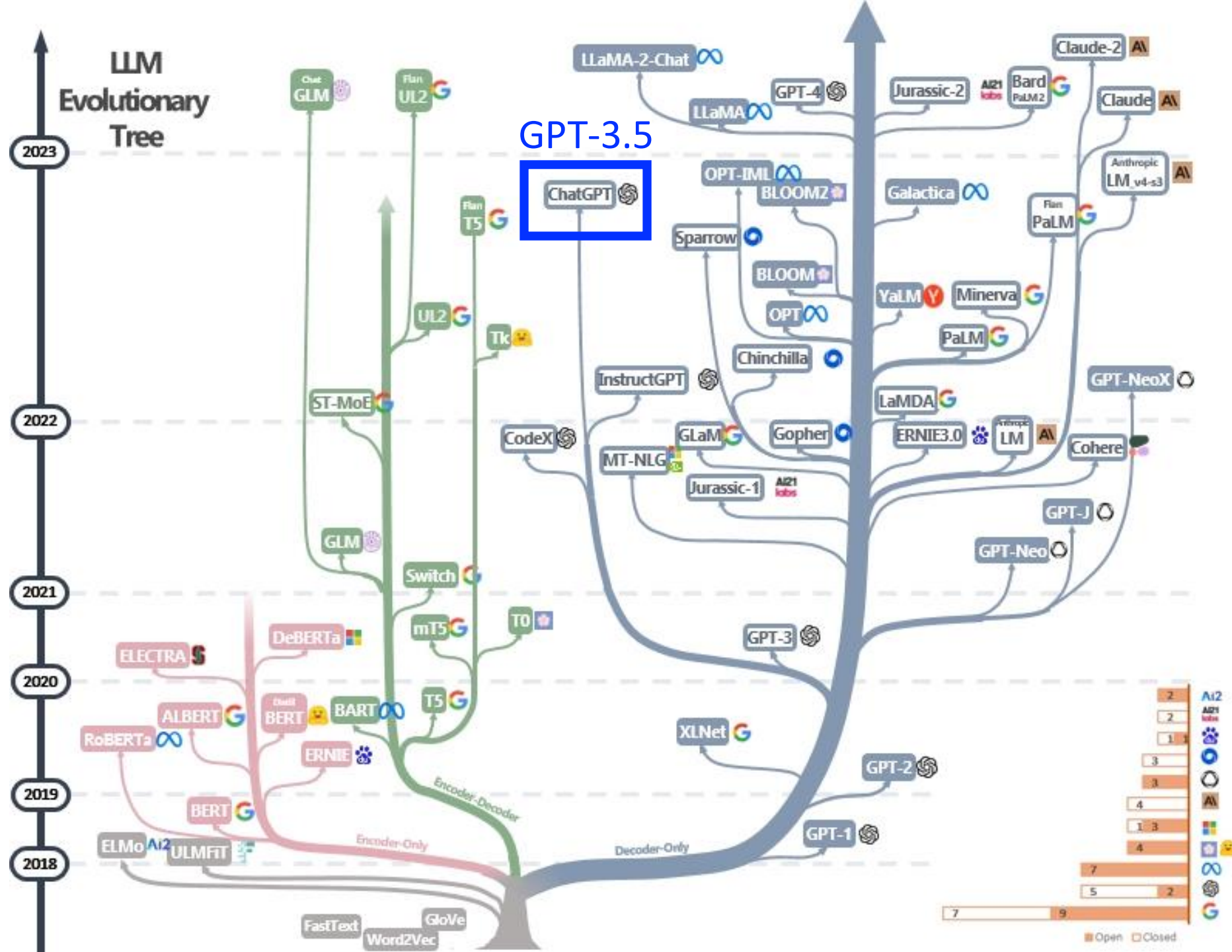
(Examples from GPT-3)

These Findings Helped Inspire a New Era of Foundation Models and Prompts

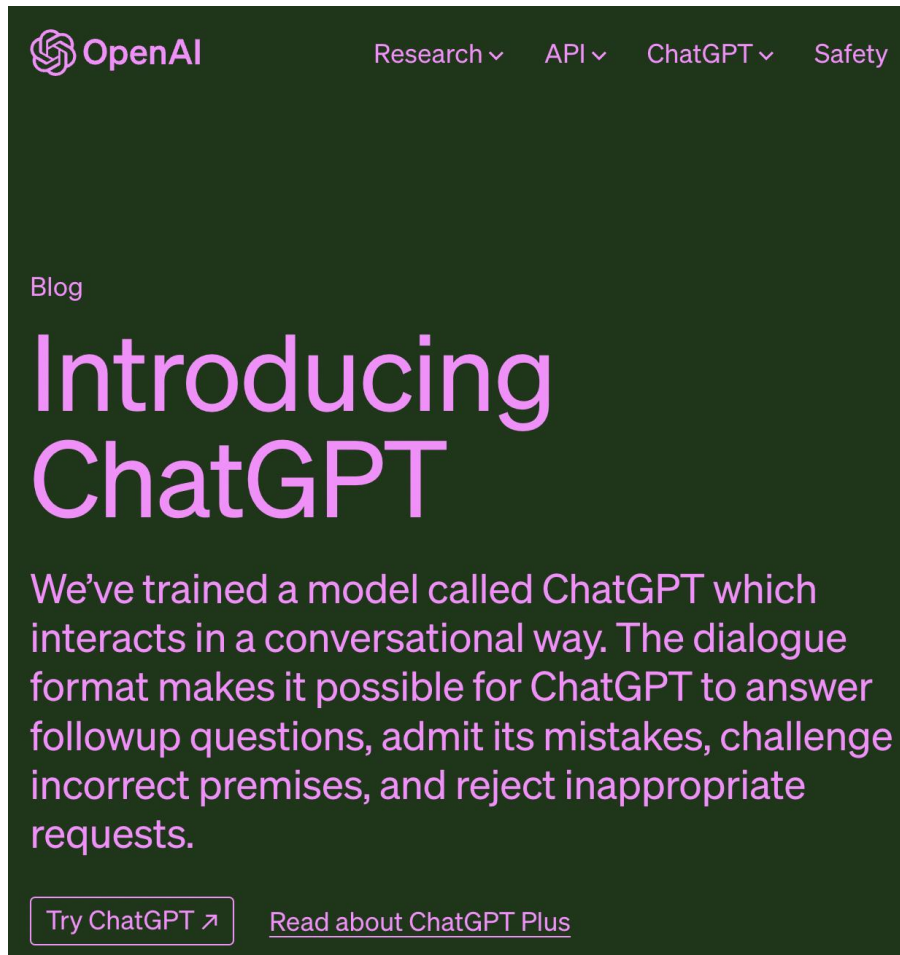
- **In-context learning**: model completes document in format modeled by task-specific examples
- **Chain-of-thought (CoT) prompting**: model conveys its reasoning process when completing the task

e.g.,

<https://github.com/Moolero0410/LLMSPracticalGuide>



November 2022: The Release of ChatGPT Changed the World's Expectations from AI...



<https://openai.com/blog/chatgpt>



And helped NVIDIA, OpenAI's GPU provider, get much richer 😊 (market value of \$4.3 TRILLION in Jan 2025)

(Excellent Summary of Prompts; ACM Surveys 2023)

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pengfei Liu

Carnegie Mellon University
pliu3@cs.cmu.edu

Weizhe Yuan

Carnegie Mellon University
weizhey@cs.cmu.edu

Jinlan Fu

National University of Singapore
jinlanjonna@gmail.com

Zhengbao Jiang

Carnegie Mellon University
zhengbaj@cs.cmu.edu

Hiroaki Hayashi

Carnegie Mellon University
hiroakih@cs.cmu.edu

Graham Neubig

Carnegie Mellon University
gneubig@cs.cmu.edu

(Another Excellent Read; arXiv 2023)

A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert,
Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt

*Department of Computer Science
Vanderbilt University, Tennessee*

e.g., “The Persona” Pattern

- “From now on, act as a security reviewer. Pay close attention to the security details of any code that we look at. Provide outputs that a security reviewer would regard regarding the code.”
- “You are going to pretend to be a Linux terminal for a computer that has been compromised by an attacker. When I type in a command, you are going to output the corresponding text that the Linux terminal would produce.”

To Help With Prompt Engineering, Many Companies Now Sell Prompts; e.g.,

The image displays two overlapping website screenshots. The background is the PromptBase website, which has a dark theme and a navigation bar with links for Marketplace, Generate, Hire, Login, and Sell. A search bar for 'Search Prompts' is visible. The foreground is the PromptAttack website, which has a white header with navigation links for ChatGPT, Midjourney, Marketing, Shop, Other, Free Prompt, and Free Generator. It also features a search bar for 'Search Prompts, @authors or #tags' and links for Marketplace, Login, and Register. The main content area of PromptAttack has a dark blue background with the text 'Unleash the power of Artificial Intelligence' and 'Prompt Attack Your #1 Prompt Marketplace'. Below this, it states 'PromptAttack is a marketplace where you can purchase and sell high-quality prompts that generate optimal stunning results while also reducing your API expenses.' There are two buttons at the bottom: 'Find a prompt' and 'Sign Up'. Three prompt cards are displayed: 'Jagged Cut Out Punk Posters' by @midrun for \$2.99, 'Research Paper Summarizer' by @laxman1986 for \$2.99, and 'Studio Quality Product Fruit...' by @midrun for \$2.99. The PromptAttack logo is a stylized blue and purple icon.

PromptBase Search Prompts Marketplace Generate Hire Login Sell

AI Models

Midjourney Marketing Shop Other Free Prompt Free Generator login

PromptAttack Search Prompts, @authors or #tags Marketplace Login Register

Unleash the power of Artificial Intelligence

Prompt Attack Your #1 Prompt Marketplace

PromptAttack is a marketplace where you can purchase and sell high-quality prompts that generate optimal stunning results while also reducing your API expenses.

Find a prompt Sign Up

Jagged Cut Out Punk Posters @midrun \$2.99

Research Paper Summarizer @laxman1986 \$2.99

Studio Quality Product Fruit... @midrun \$2.99

Prompt Engineering Work Also Helped Inspire Rise of “Reasoning” Models; e.g.,

- Sep 2024: OpenAI’s O1, world’s first reasoning model
- Nov 2024: Alibaba’s QWQ, or Questions with Qwen (*model open-source; extends Meta’s Llama*)
- Dec 2024: Google’s Gemini Flash Thinking and OpenAI’s O3
- Jan 2025: DeepSeek’s R1 (*model open-source*)

To Help With Leveraging These Huge Models, Companies Now Sell Access to Them; e.g.,

together.ai

Products ▾ For Business ▾ For Developers ▾ Pricing ▾ Research Company ▾ Docs Contact Get Started ↗

Hyperbolic

About Docs Blog Careers Pricing Schedule a Call Launch App

Azure AI model catalog

Discover, evaluate, customize, and deploy AI models

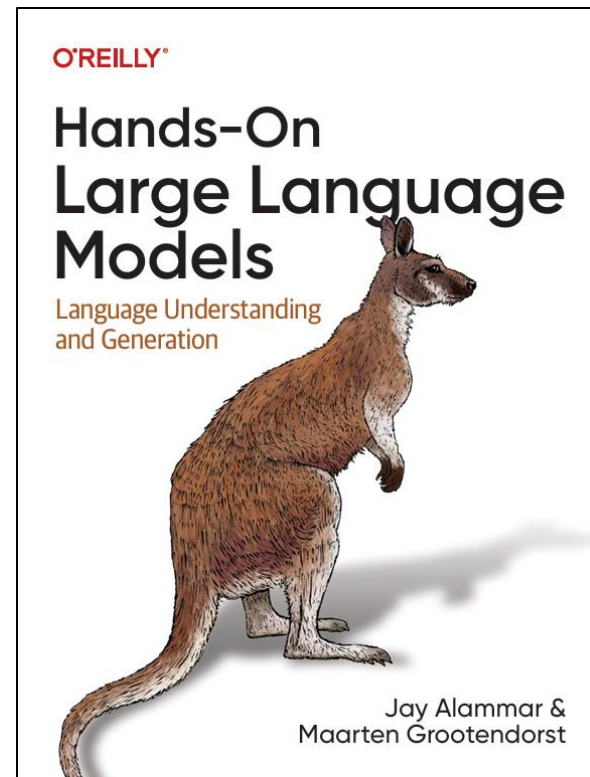
 Innovate faster with popular models from creators such as Microsoft, OpenAI, Mistral, Meta, Stability AI, Core42, and Nixtla.

[Get started with Azure](#)

Th

Programming Tutorials

- Everley's programming tutorial for Hugging Face Inference Playground
- Ch 6 of this book:



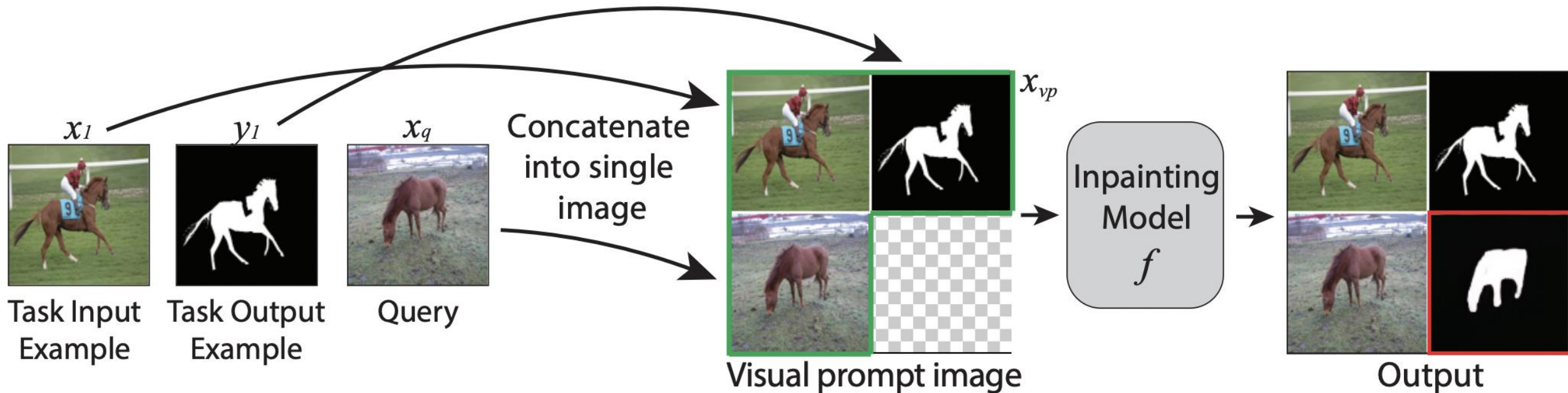
Today's Topics

- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

Motivating Observation

- Foundation models achieve better performance for NLP tasks when provided “in-context” examples.
 - i.e., [Task description, Examples, Prompt]
 - e.g., “Translate English to Spanish. Computer -> Computadora. Vision ->
- Idea: Use in-context few-shot learning for image-based prompts

Novel Idea: Image Inpainting



Designed to adapt to any “image-to-image translation” task by using the model as is (e.g., no fine-tuning required)

Approach: Prompting for Image Inpainting



Edge detection



Colorization



Inpainting



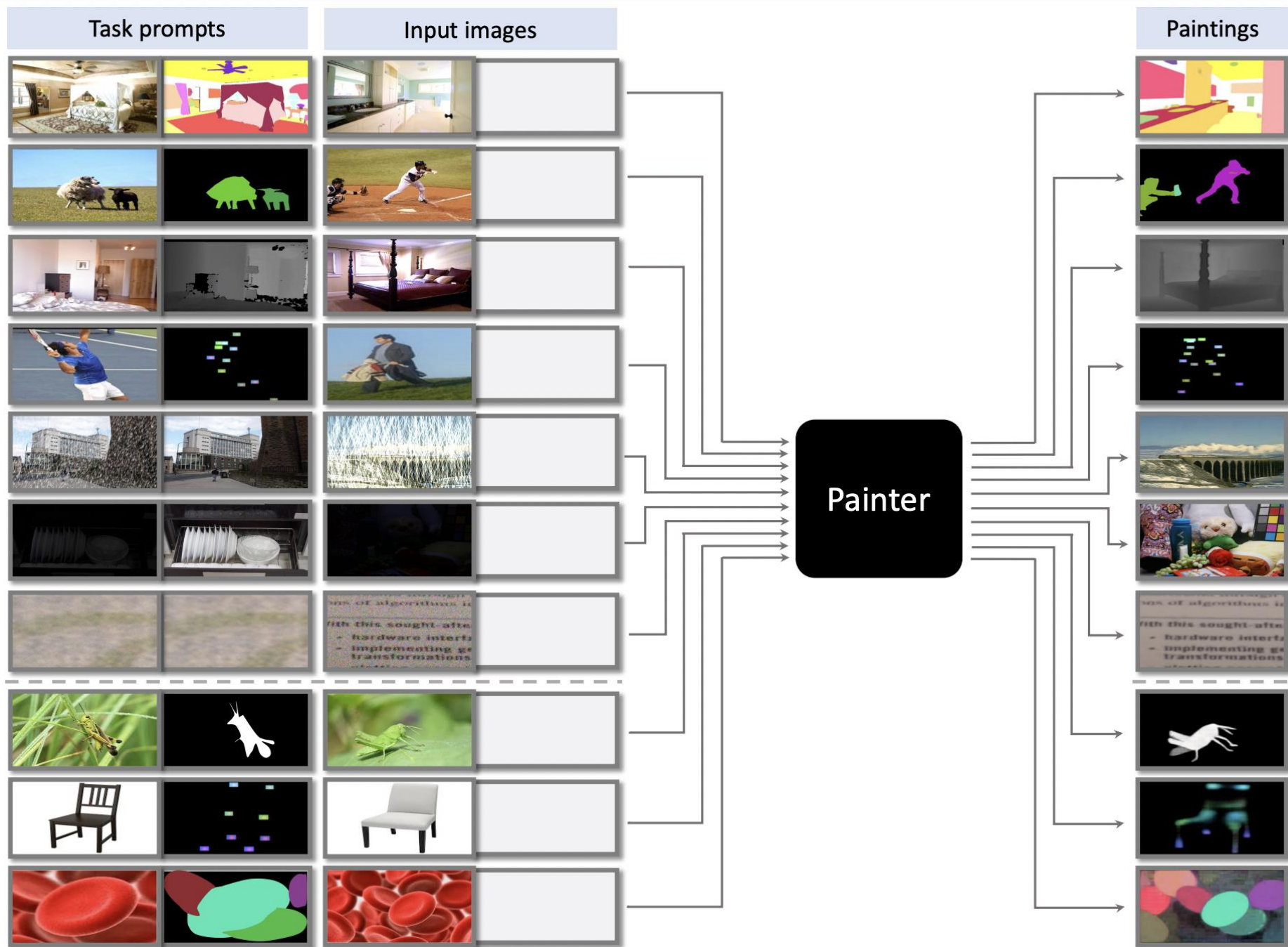
Segmentation



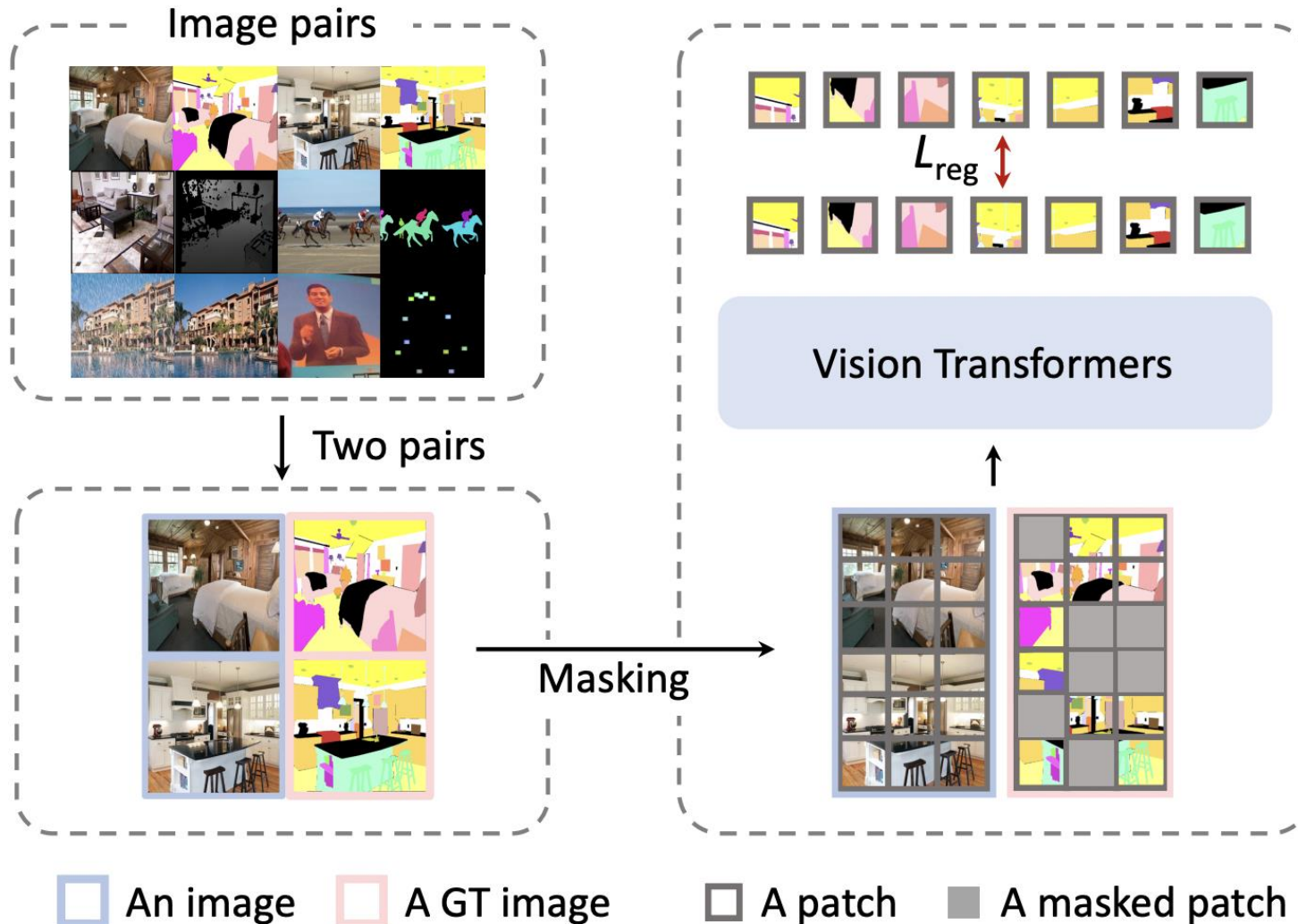
Style transfer

Approach

Extended in 2023
to standard vision
benchmarks:



Training: Masked Image Modeling



Uses self-supervised learning such that the model predict values in masked out patches

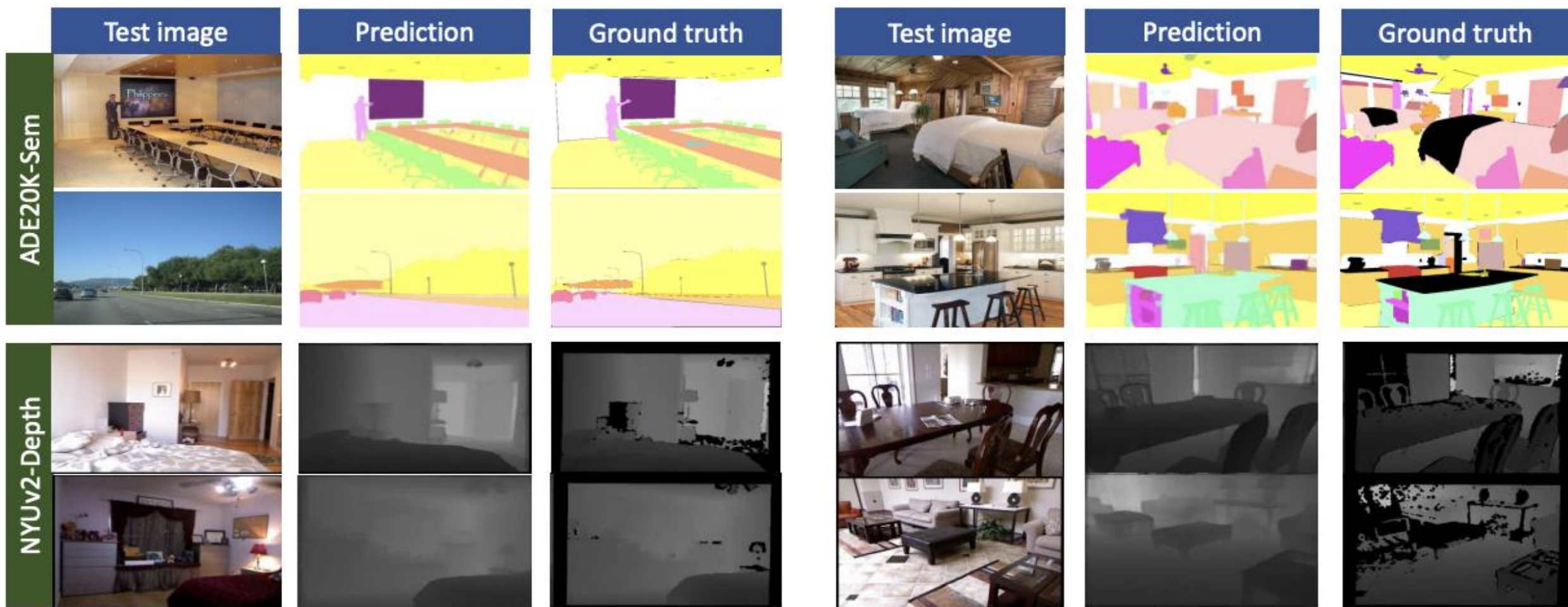
Uses standard vision benchmarks for each evaluated task

Experimental Results

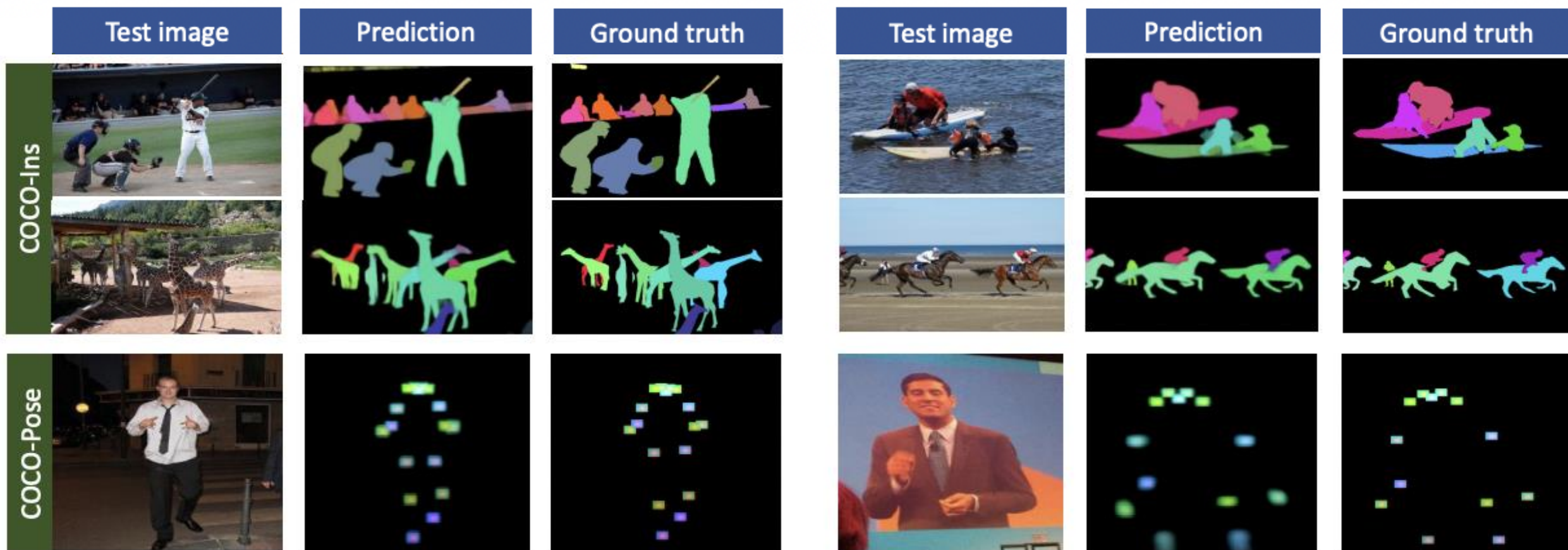
(Used in prompt the best performing example-per pair per task from all examples in the training dataset)

Model achieves state-of-the-art performance on depth estimation for NYUv2 dataset and outperforms other generalist models on several more tasks.

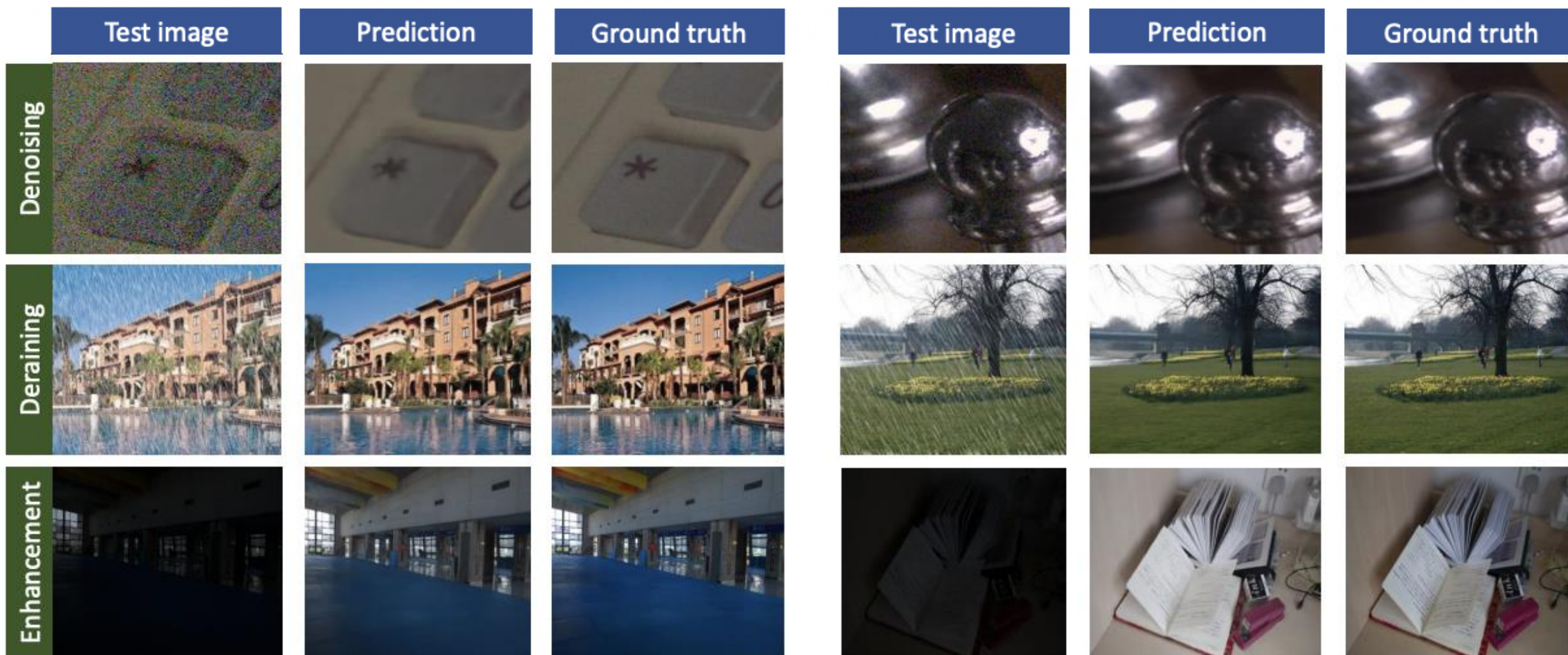
Qualitative Results: In-Domain Results



Qualitative Results: In-Domain Results

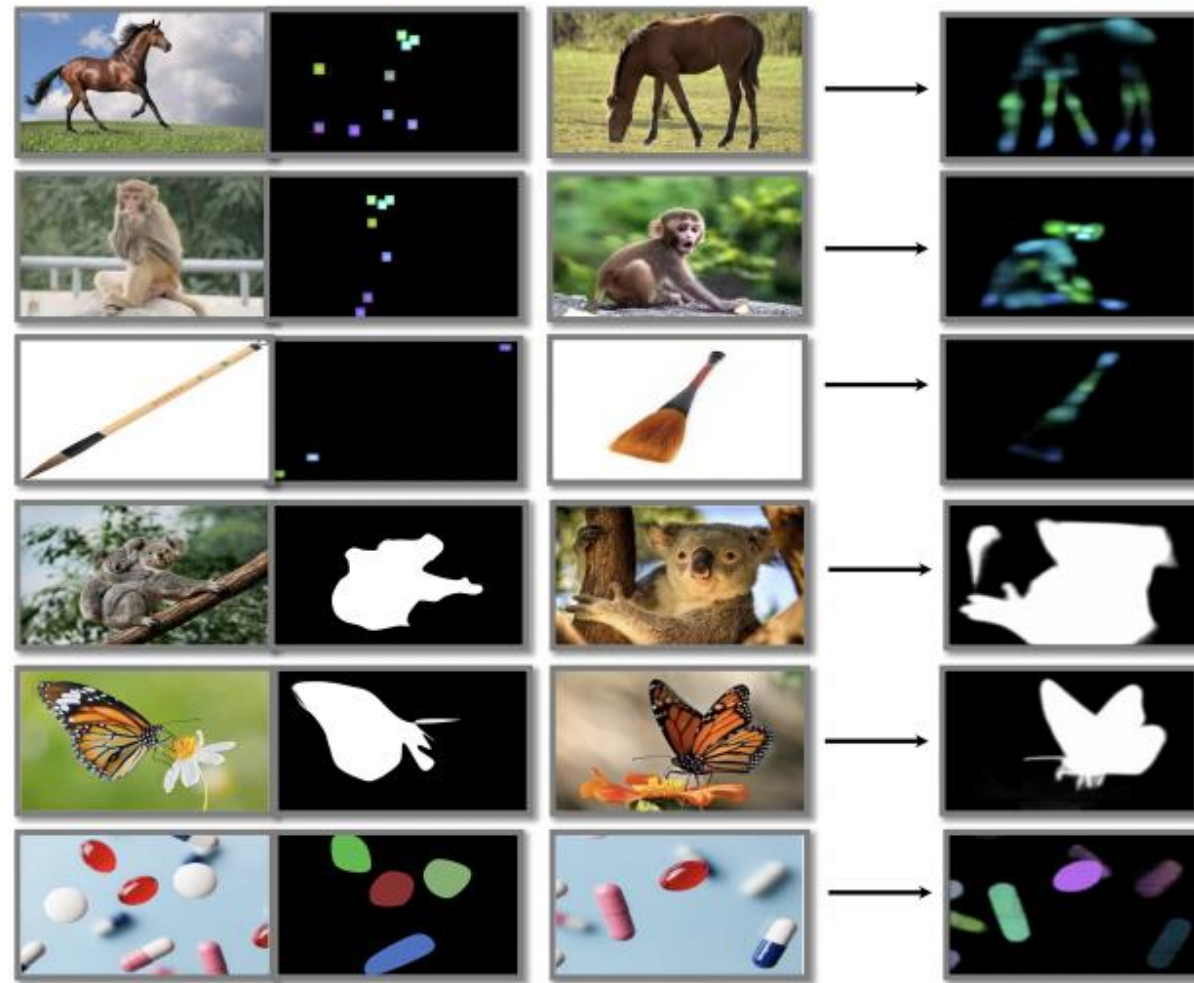


Qualitative Results: In-Domain Results



Qualitative Results: Open-Vocabulary Results (i.e., Categories Not Seen at Training)

Shows in-context examples, prompts, and predictions for keypoint detection, object segmentation, and instance segmentation



Today's Topics

- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”

(Pioneering LMM; Neurips 2022)

 **Flamingo: a Visual Language Model
for Few-Shot Learning**

Jean-Baptiste Alayrac^{*,‡} Jeff Donahue^{*} Pauline Luc^{*} Antoine Miech^{*}

Iain Barr[†] Yana Hasson[†] Karel Lenc[†] Arthur Mensch[†] Katie Millican[†]

Malcolm Reynolds[†] Roman Ring[†] Eliza Rutherford[†] Serkan Cabi Tengda Han

Zhitao Gong Sina Samangooei Marianne Monteiro Jacob Menick

Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh

Mikolaj Binkowski Ricardo Barreira Oriol Vinyals Andrew Zisserman

Karen Simonyan^{*,‡}

^{*} Equal contributions, ordered alphabetically, [†] Equal contributions, ordered alphabetically,

[‡] Equal senior contributions

DeepMind

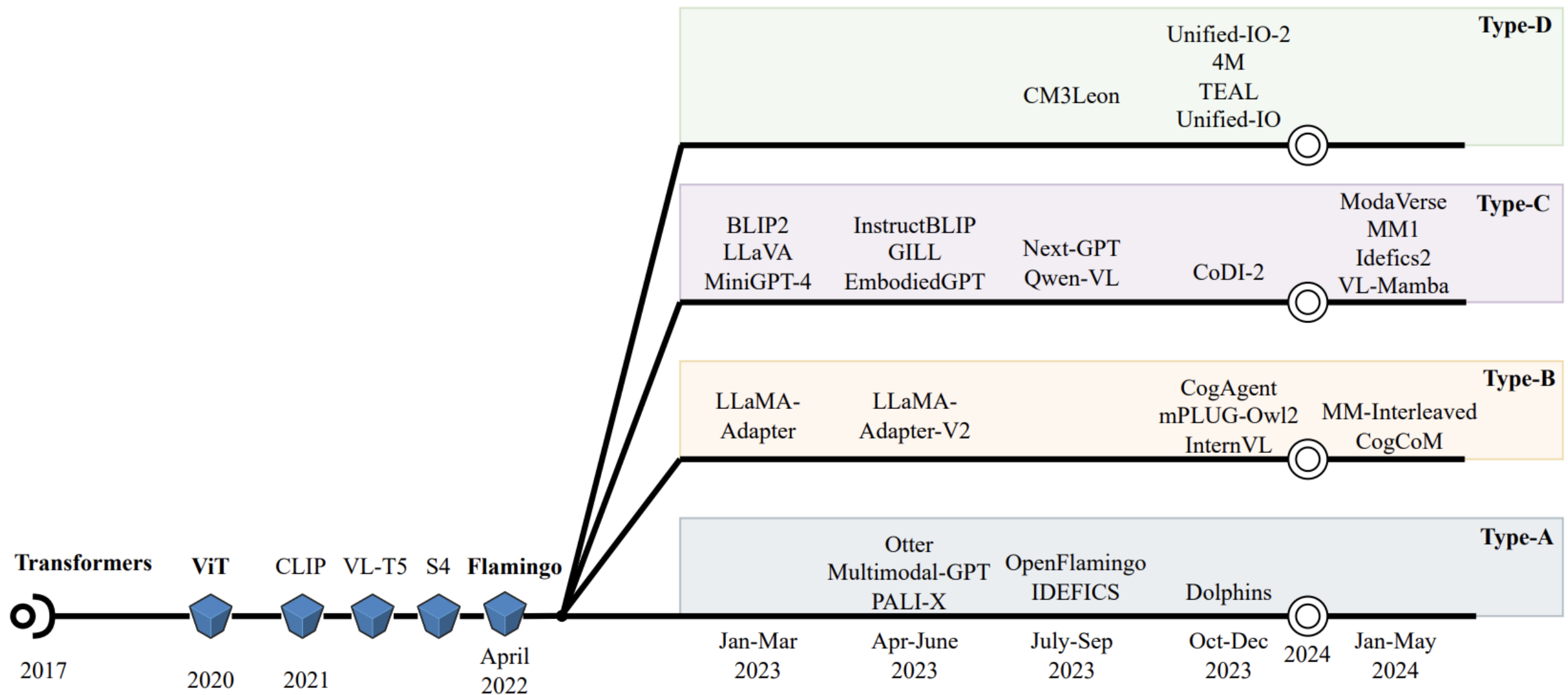
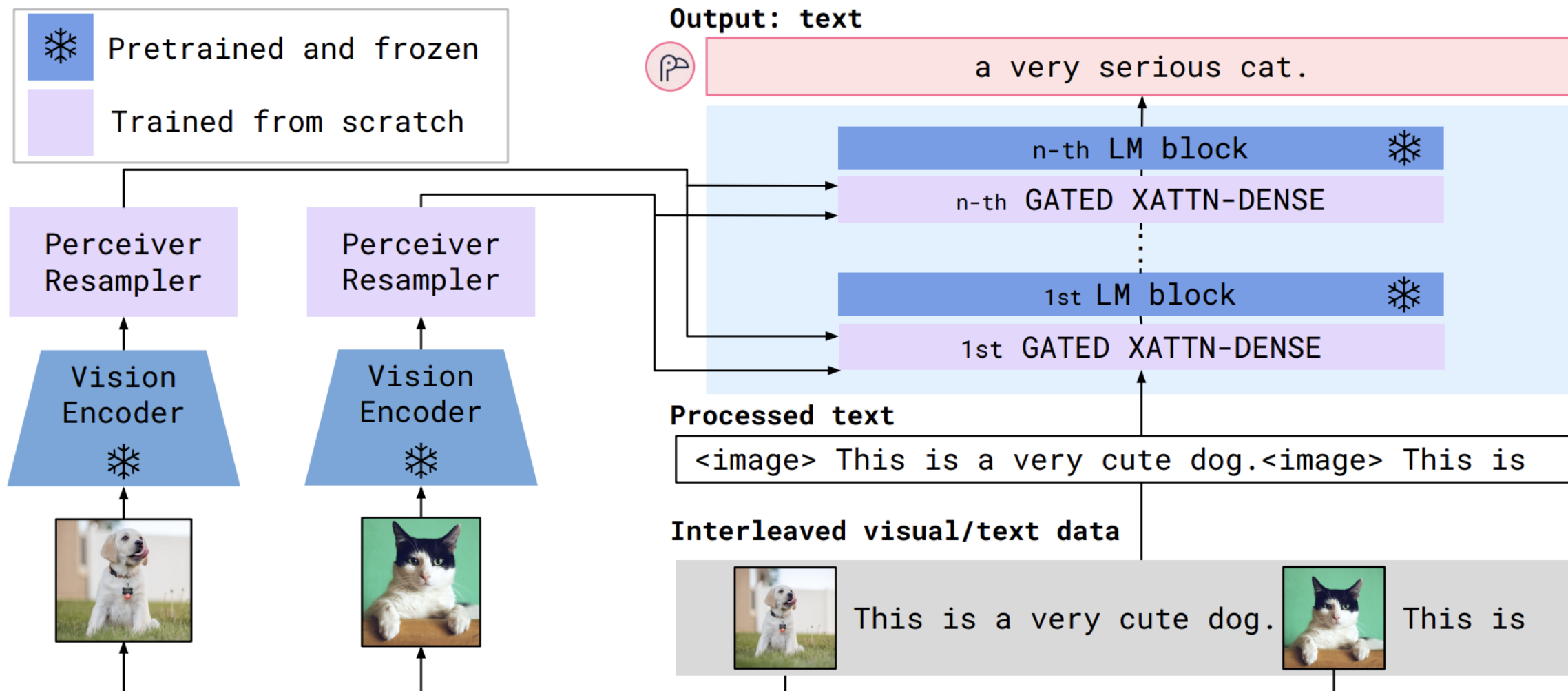
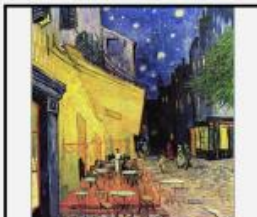


Figure 1: Development timeline of Multimodal models grouped in four proposed architecture types.

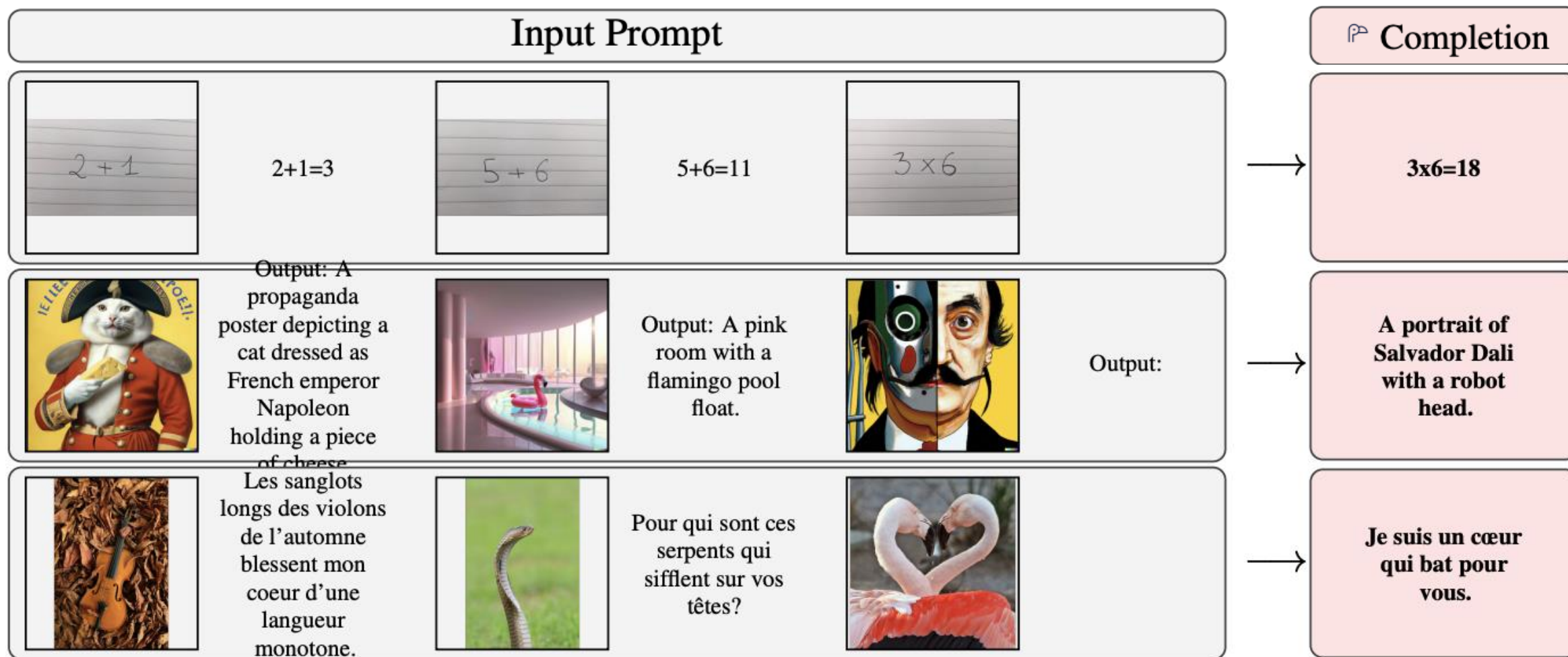
Key Challenge: Interleaved Image-Text Input



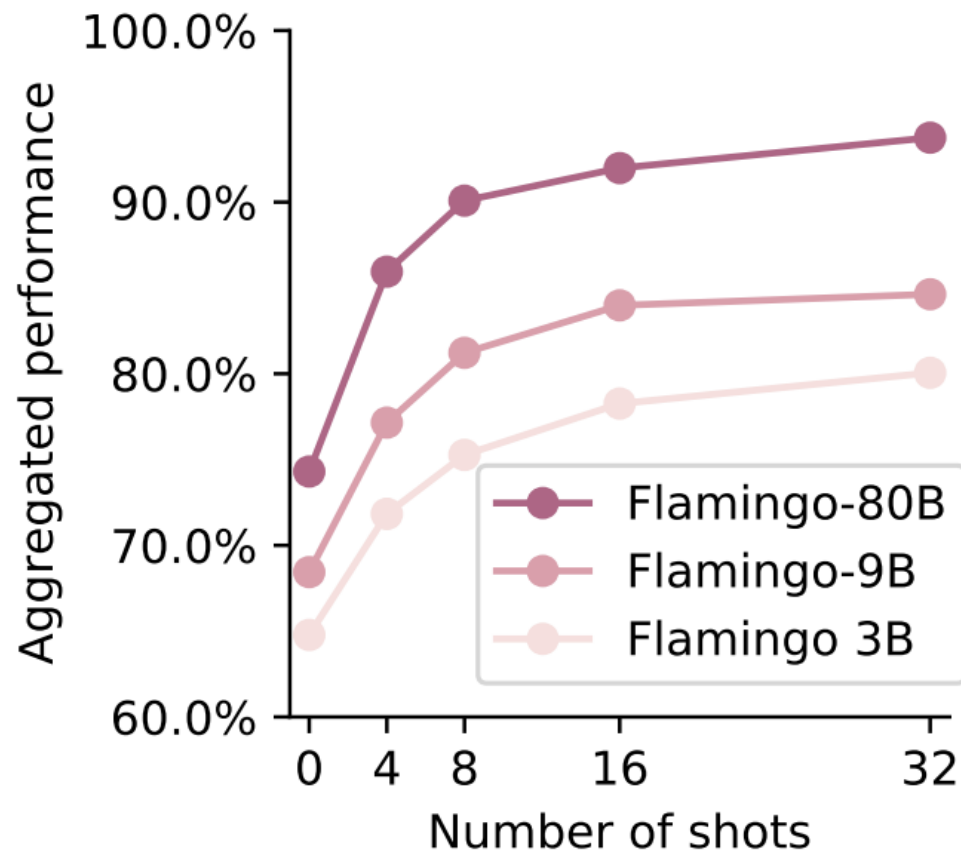
Achieved Strong Zero-Shot Performance with Multimodal Prompts for Many Tasks; e.g.,

Input Prompt					Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.	 This is	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.	 What is the name of the city where this was painted? Answer:	Arles.
	Output: "Underground"		Output: "Congress"	 Output:	"Soulomes"

Achieved Strong Zero-Shot Performance with Multimodal Prompts for Many Tasks; e.g.,



Results: Larger Models Perform Better



What are trends for different model sizes and numbers of “shots”?

Largest model achieved [state-of-the-art](#) on 6 of 16 tasks, outperforming fine-tuned models

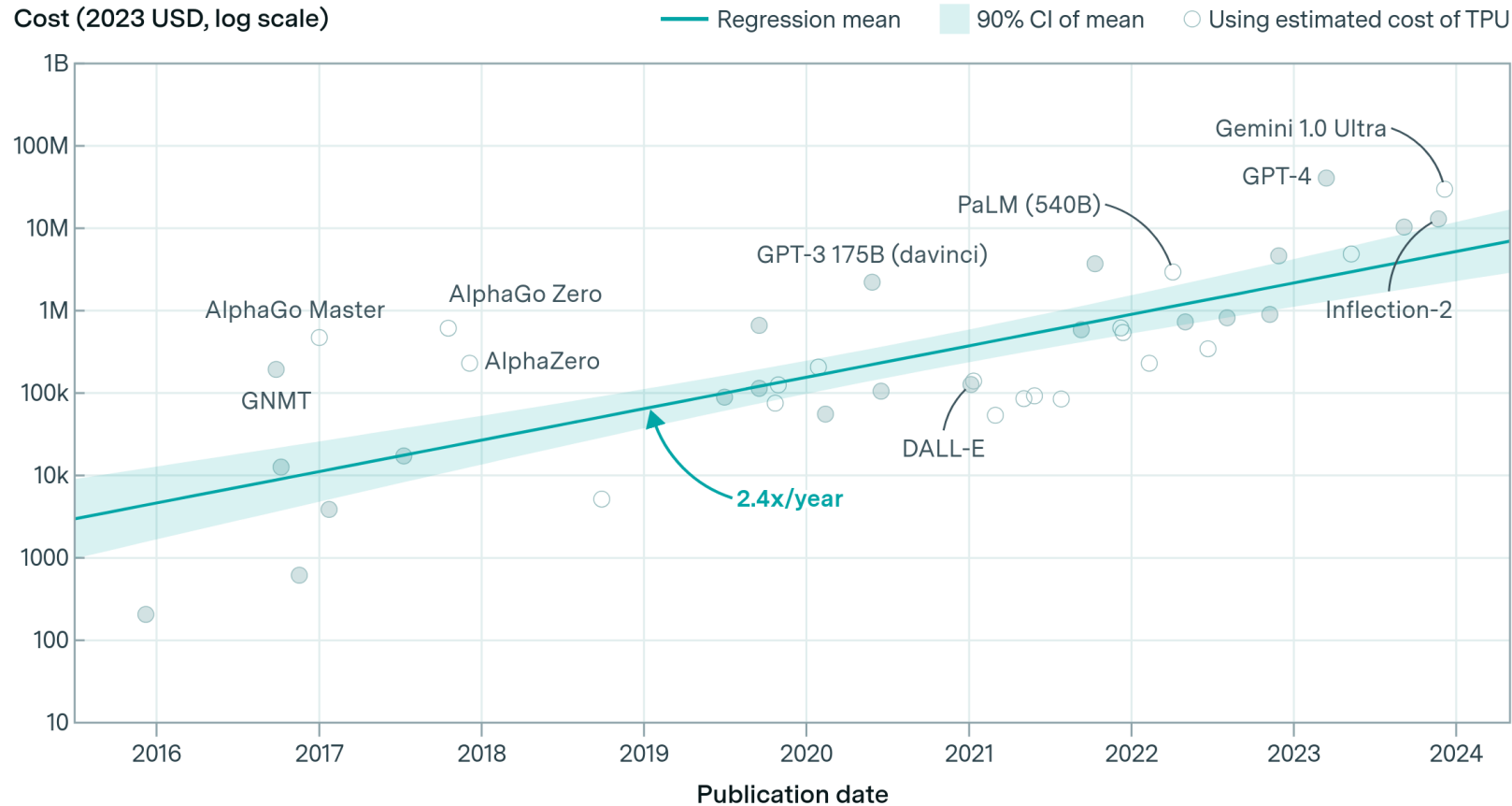
LMMs Extending Beyond Language + Vision

- e.g., PandaGPT: text, image, video, audio, and sensors for depth (3D), thermal (infrared radiation), or inertial measurement units (IMU) of motion and position; connects objects with sound, 3D shape, temperature, and motion
- e.g., SpeechGPT: language and speech
- e.g., NExT-GPT: text, images, videos, and audio

Class Discussion: What are Risks of Using Foundation Models (Today, Generative AI)?

- Model biases/limitations can affect all downstream models

Scaling Laws Still In Play: Training Costs (and Emergent Abilities) Are Soaring



“Existential” threat for those without “deep pockets”; **how to make a contribution?**

Next Two Lectures: How Can You Contribute to Neural Networks and Deep Learning in this “Scaling Law” Era?



Today's Topics

- Motivation
- Foundation Models
- Prompting “Large Language Models” for NLP
- Prompting “Large Vision Models” for Computer Vision
- Prompting “Large Multimodal Models”



The End