Foundation Models and Prompts

Danna Gurari
University of Colorado Boulder
Spring 2024

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/

Review

- Last lecture:
  - Explosion of transformers
  - GPT
  - BERT
  - ViT
  - Programming tutorial
- Assignments (Canvas):
  - Problem set 4 due Wednesday
  - Project outline due after Spring break
- Questions?

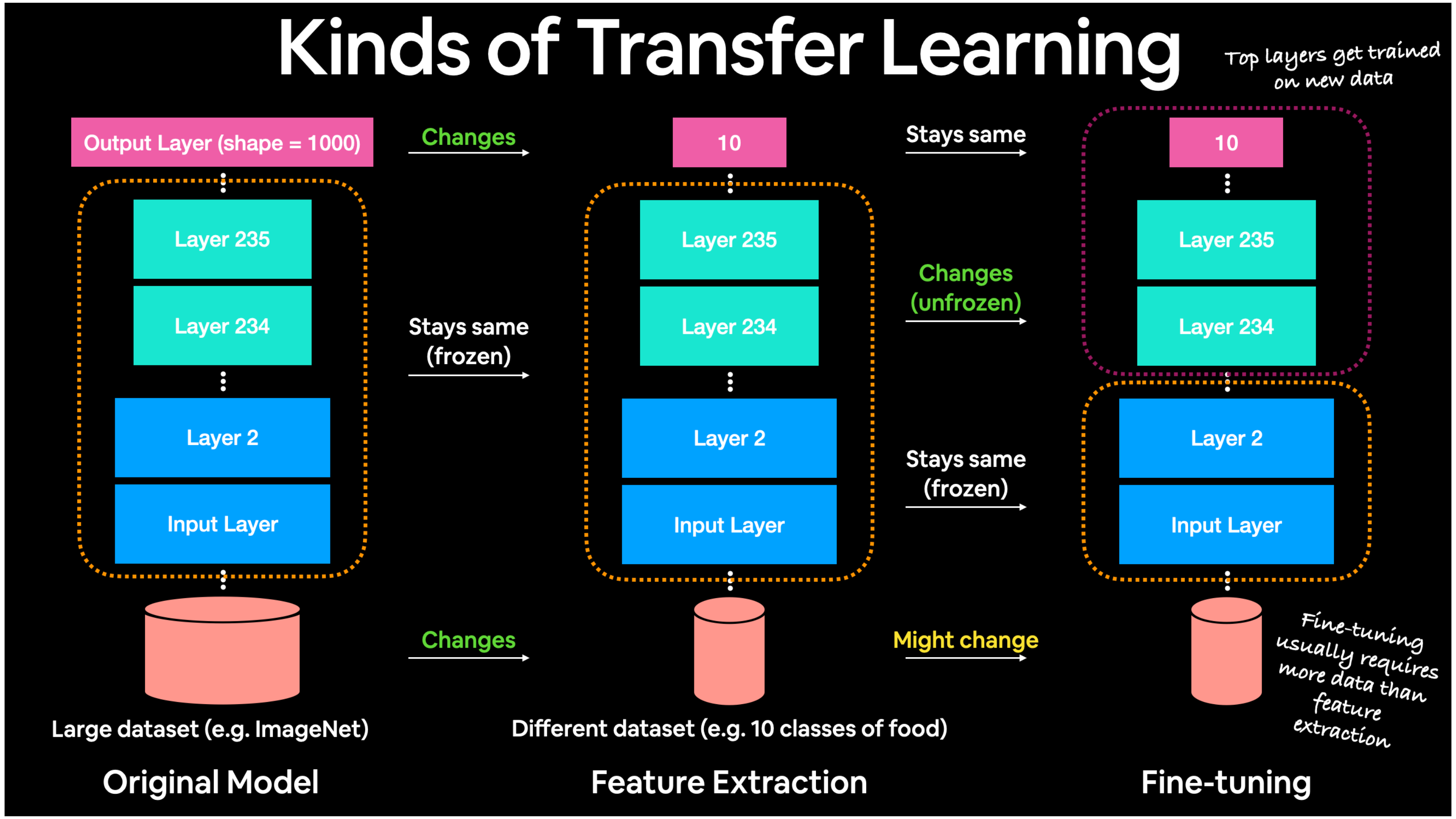
Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

What We Learned Works Over Past Decade

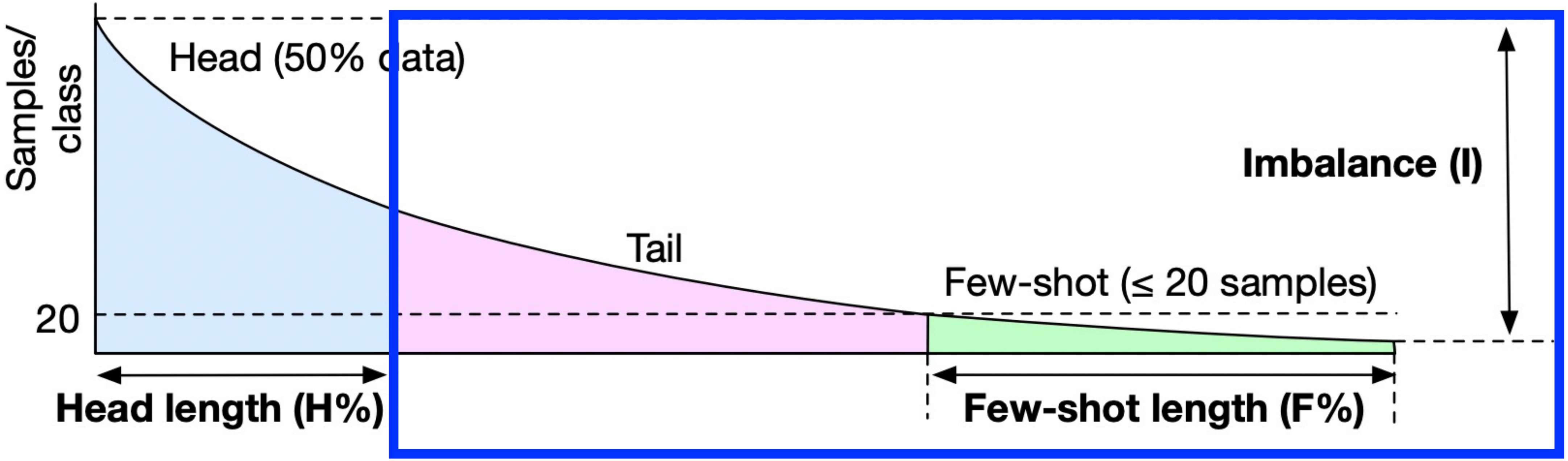


Can achieve strong performance with lots of labeled data for target task (aka closed world problems) when training from scratch or fine-tuning

https://dev.mrdbourke.com/tensorflow-deep-learning/04_transfer_learning_in_tensorflow_part_1_feature_extraction/

Open Problems: Beyond Big Data

Learning with Limited Labeled Training Data
(e.g., Few-Shot Learning)



Perrett et al. Use Your Head: Improving Long-Tail Video Recognition. CVPR 2023.

Open Problems: Beyond Closed-World Setting

Vocabulary / Language

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

Visual Content

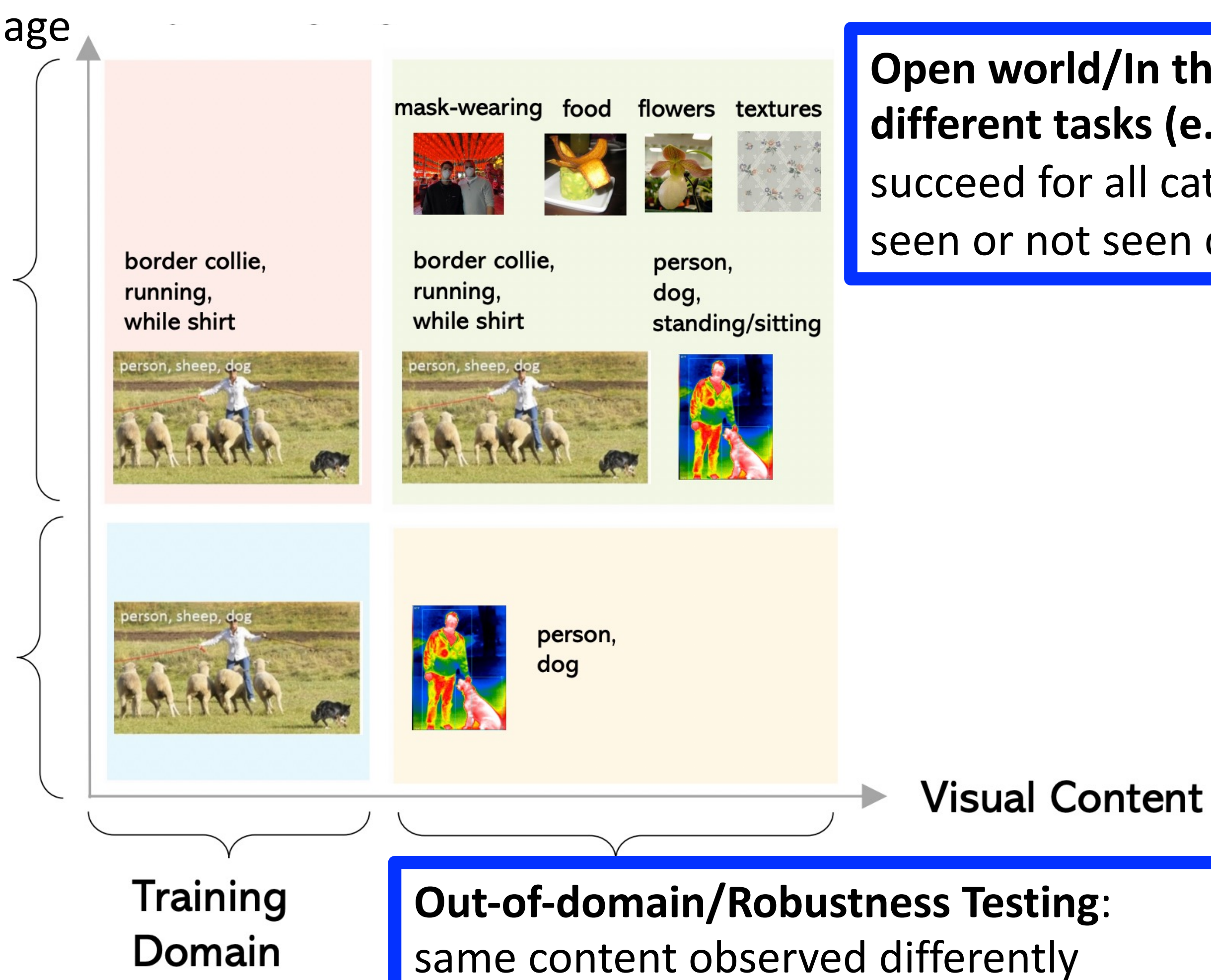
Training Domain

https://arxiv.org/pdf/2210.09263.pdf

Open Problems: Beyond Closed-World Setting

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

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

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

https://arxiv.org/pdf/2204.08790.pdf

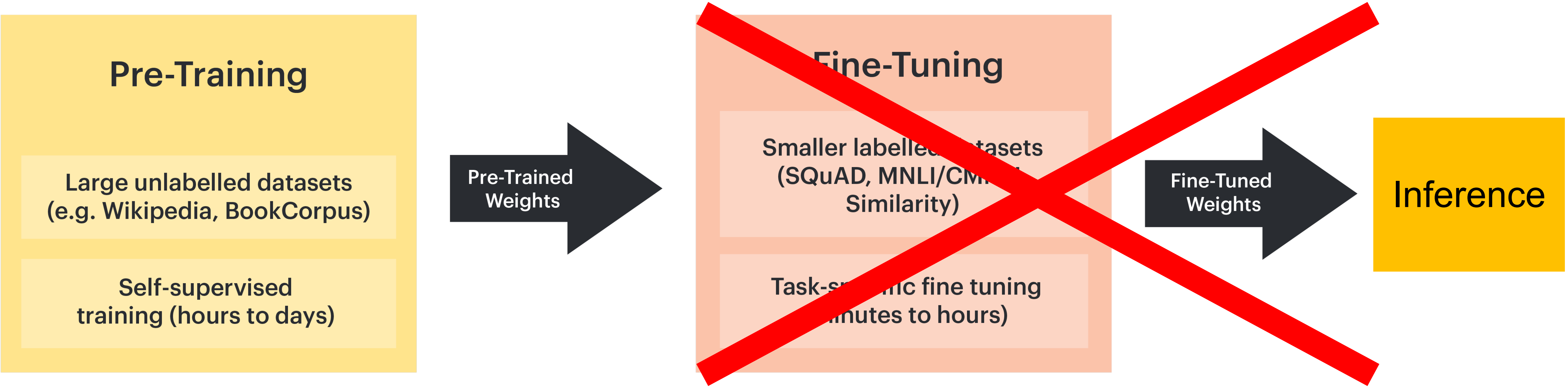
New Paradigm:

Current Findings Suggest Foundation Models Generalize Well With Limited Training Data and Beyond Closed World Tasks

Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

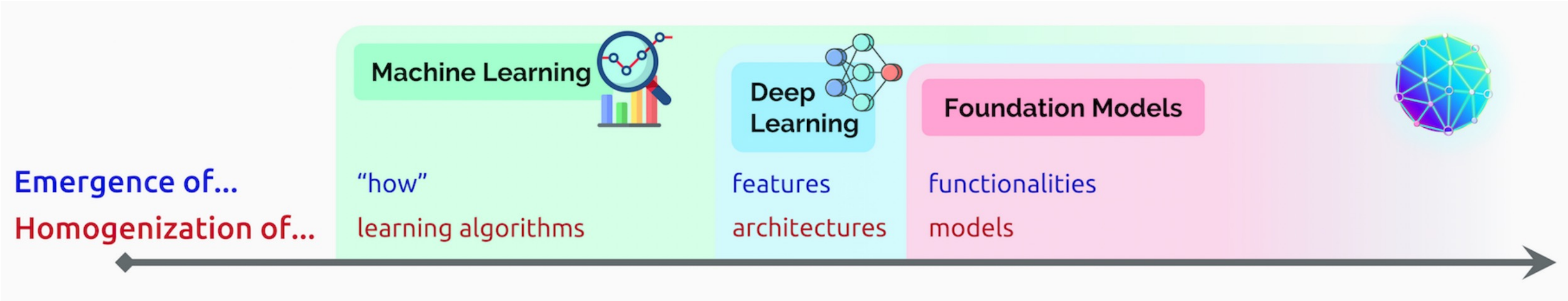
Foundation Model: Key Idea



Emergent behavior discovered around 2018 (in NLP) that a foundation model can be used ***as is*** for many downstream tasks with ***prompting!***

https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

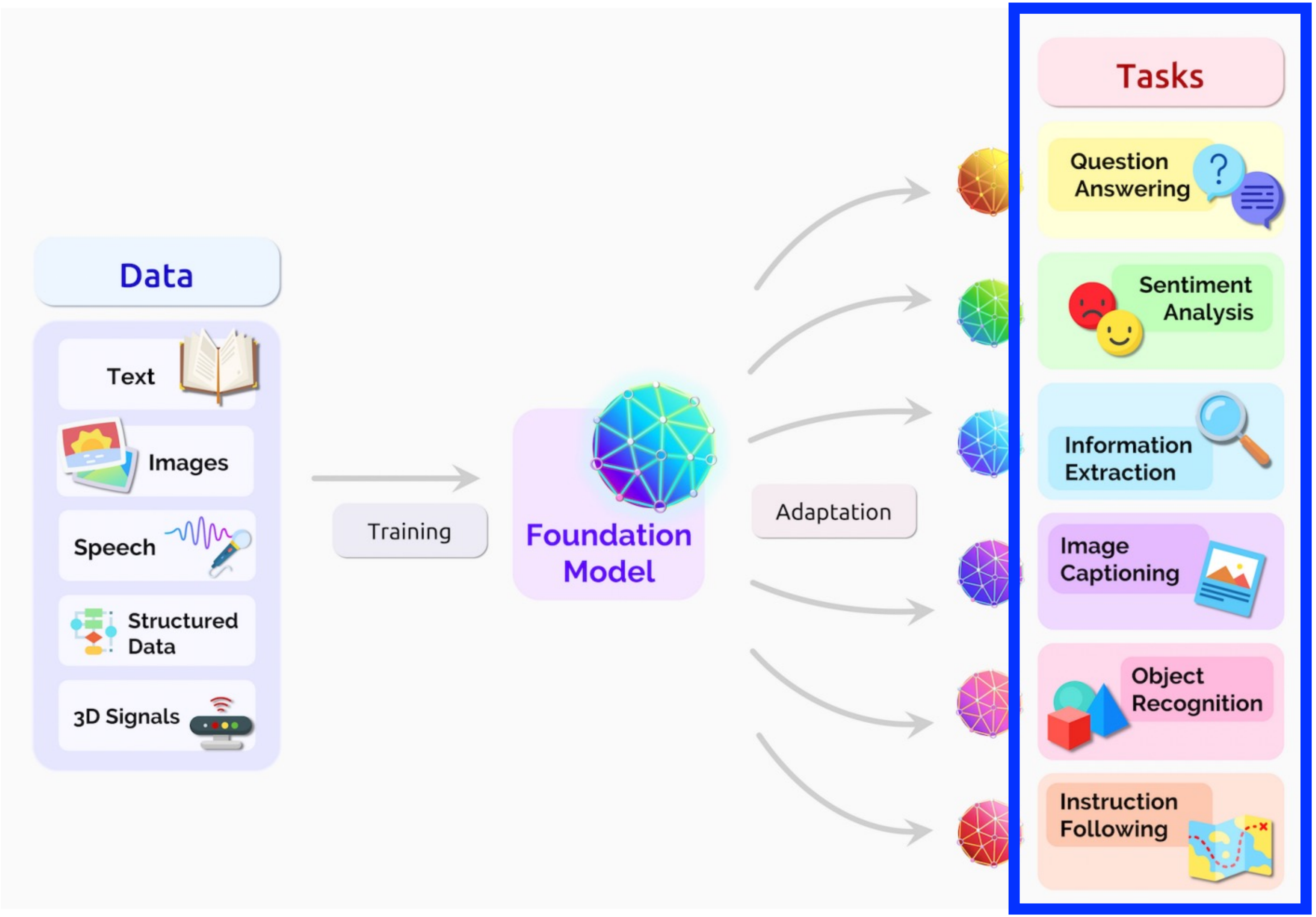
Definition of “Foundation Model”



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

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Foundation Models: Training to Evaluation



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 versus data from blind people)

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Foundation Models: Why Now?

Key ingredients identified:

1. Transformer model architecture

2. Lots more training data by using Internet data

3. Sufficient hardware with modern GPUs

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Caution: Risks of Using Foundation Models

- e.g.,
  - Model biases/limitations can affect all downstream models
  - Computationally expensive models (and so bad for environment)

Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

(Pioneering Paper for Foundation Models; 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**

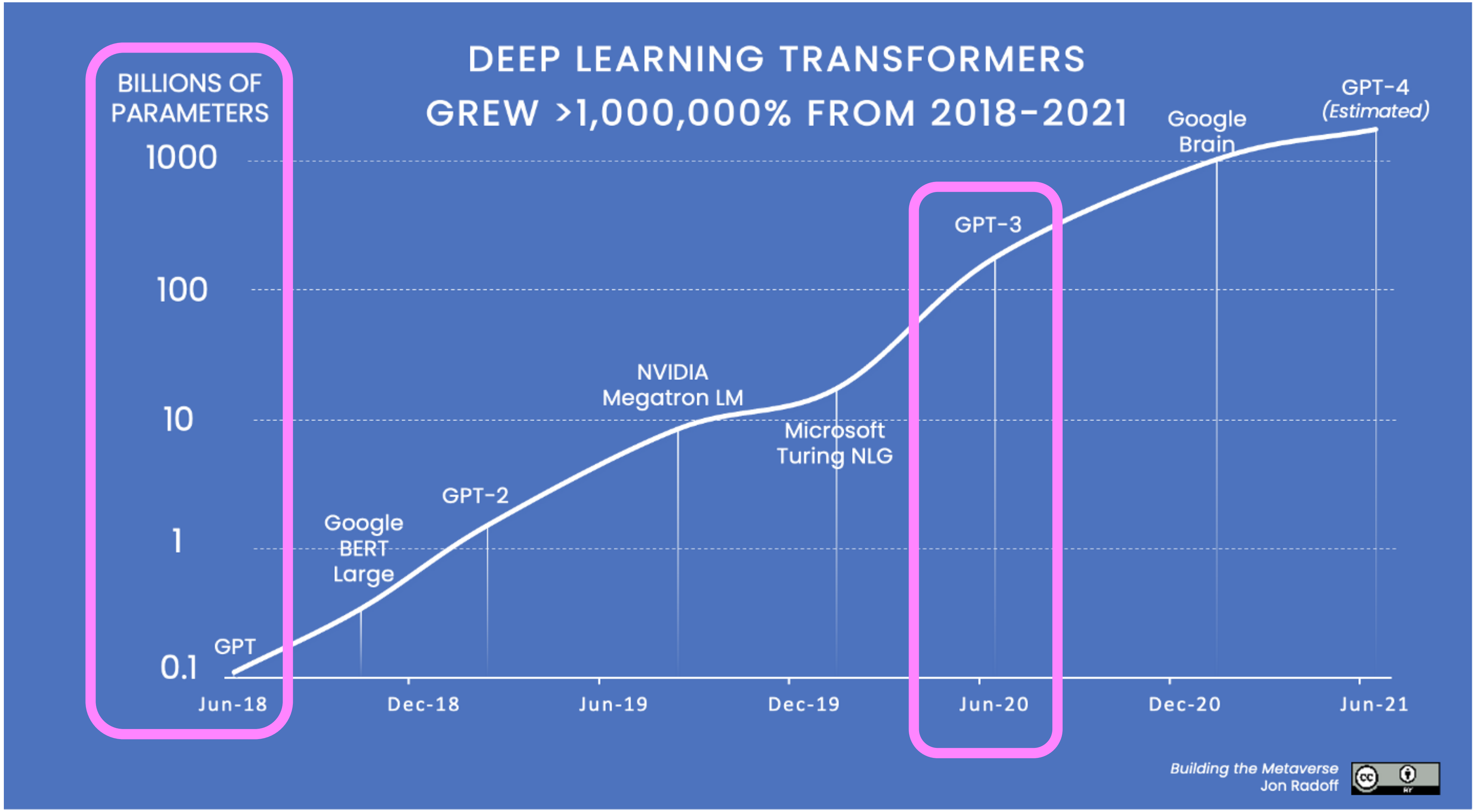
Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Key Idea: Increase Model Size

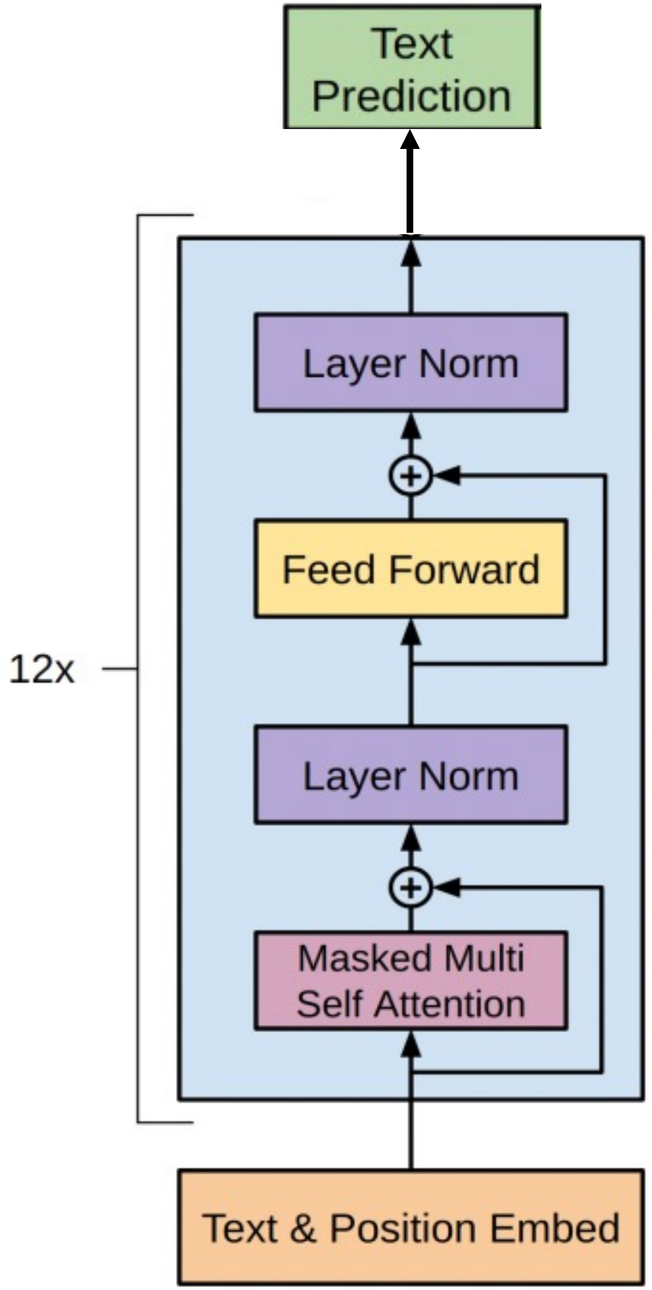


https://medium.com/building-the-metaverse/the-metaverse-and-artificial-intelligence-ai-577343895411

GPT-3: Model Design

Slightly modified version of GPT with many more layers!

Extends GPT-2 (Radford et al. OpenAI blog 2019)



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

GPT-3: Model Design (8 Tested Variants)

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

GPT-3: Training Data

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%

(Additional 4 sources are known high-quality datasets to supplement the noisy, lower quality Common Crawl data)

Language composition: by word count, 93% English

Common Crawl (popular NSP source)
(web archives without HTML markup and non-text content https://commoncrawl.org/)

Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

GPT-3: Prompts Include Instructions &, Optionally, 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.

`Translate English to French:`	*task description*
`cheese =>`	*prompt*

One-shot

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

`Translate English to French:`	*task description*
`sea otter => loutre de mer`	*example*
`cheese =>`	*prompt*

Few-shot

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

`Translate English to French:`	*task description*
`sea otter => loutre de mer`	*examples*
`peppermint => menthe poivrée`	
`plush girafe => girafe peluche`	
`cheese =>`	*prompt*

Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

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 langauge tasks, the format for zero-shot learning is "Q: What is the {language} translation of {sentence} A: {translation}."

Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

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.
```

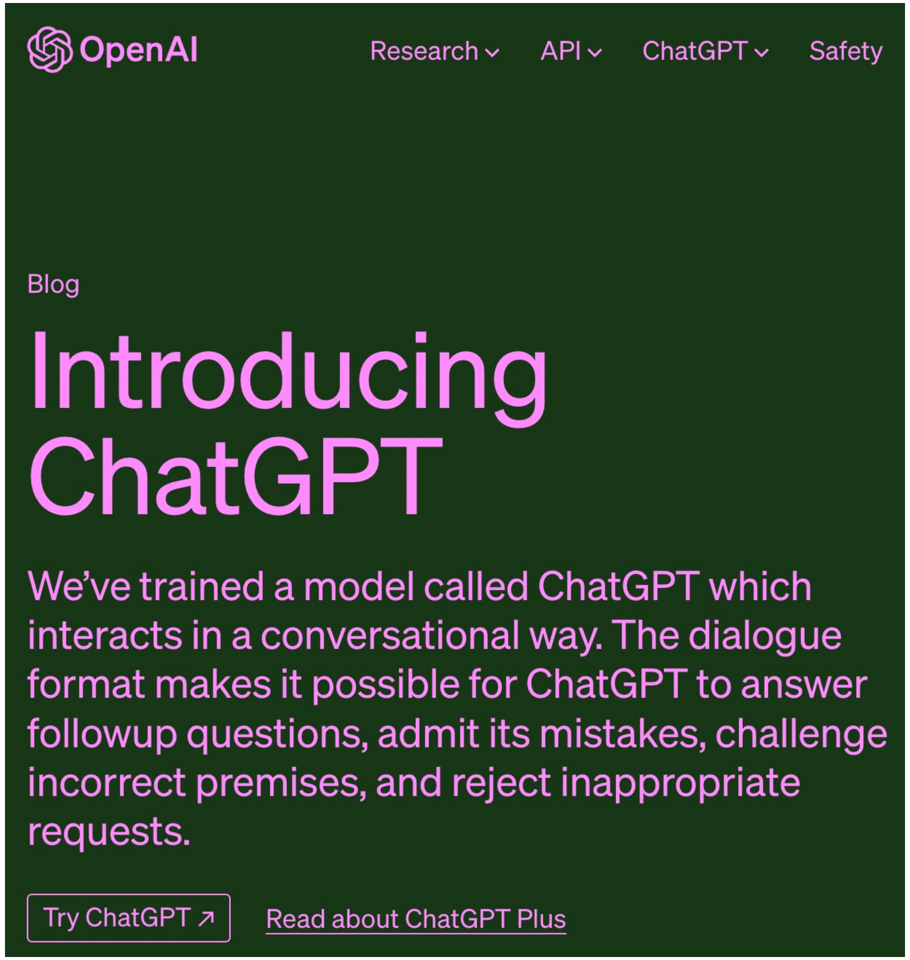
Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

Experimental Findings

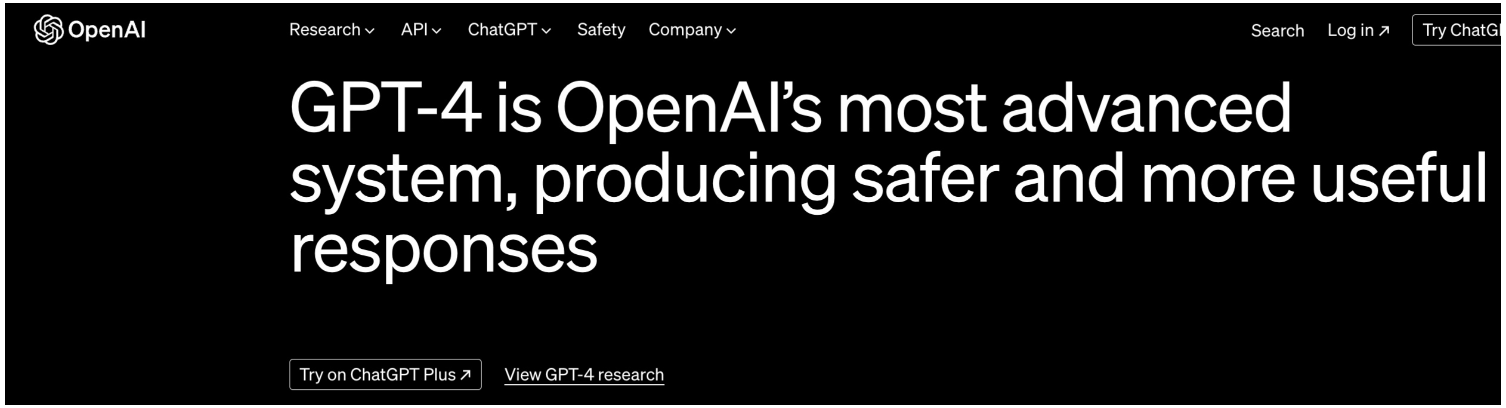
Tested on 10s of NLP datasets, showing strong performance overall and occasionally state-of-the-art performance!

Brown et al. Language Models are Few-Shot Learners. Neurips 2020.

Successors: GPT 3.5 and GPT 4



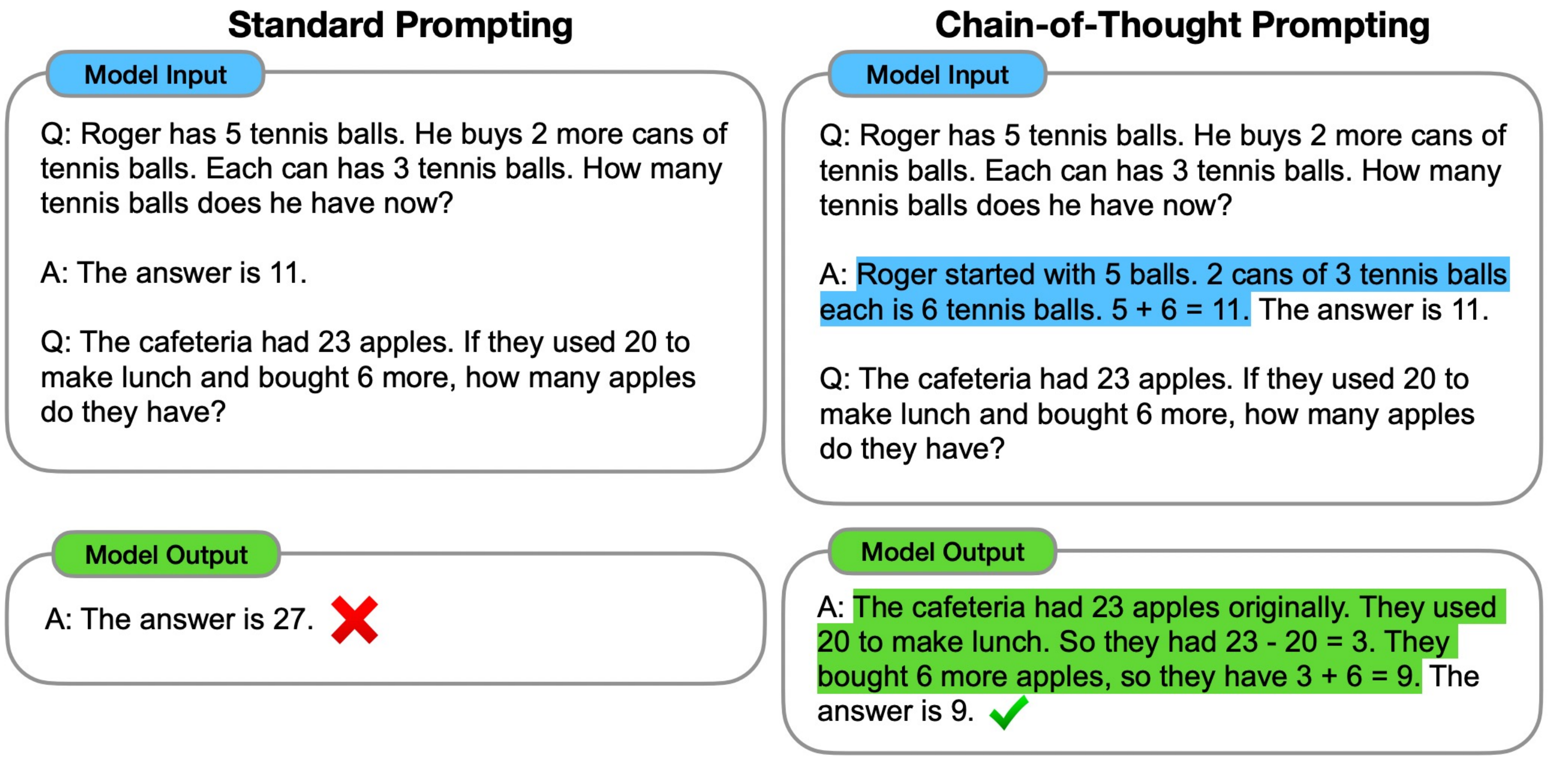
https://openai.com/blog/chatgpt



https://openai.com/gpt-4

Challenge: What Prompts to Use?

- **Chain-of-thought** prompting can help by guiding model to show its intermediate reasoning steps!



Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Neurips 2022.

Challenge: What Prompts to Use?

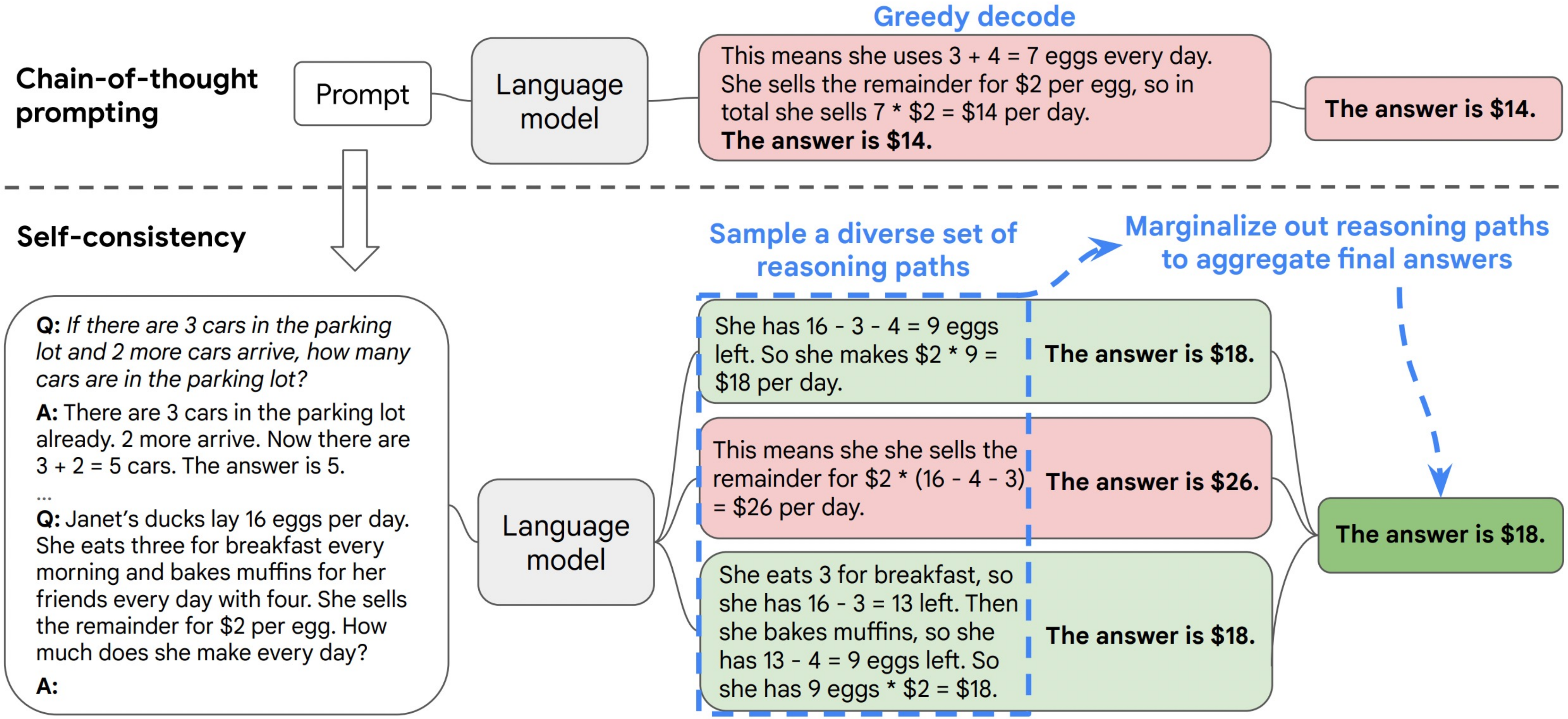
- Why COT prompting works? examples may reveal the target output format as performance still improves with invalid examples; e.g.,

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. ✓
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. ✓

Wang et al. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. arXiv 2022.

Challenge: What Prompts to Use?

Uses most consistent answer from multiple predicted answers! (as is, only applicable when answers are from a fixed set)



Wang et al. Self-consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023.

Challenge: What Prompts to Use?

Models can also perform better when asked to show their reasoning steps *without* seeing reasoning examples.

e.g., for GPT-3:

(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. ✓

Kojima et al. Large Language Models are Zero-Shot Reasoners. Neurips 2022.

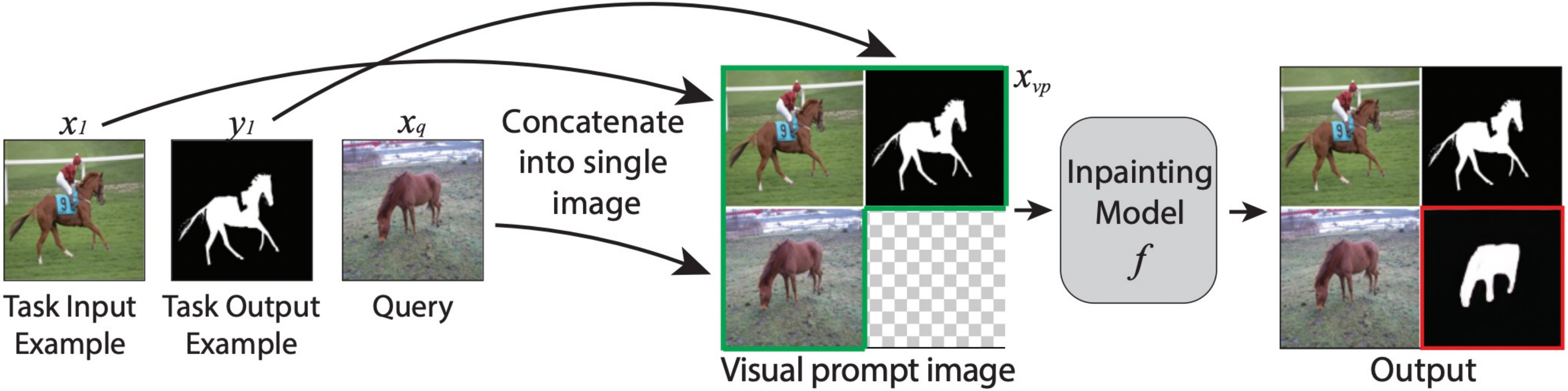
Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

Motivating Observation

- Foundation models achieved 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)

Bar et al. Visual Prompting via Image Inpainting. Neurips 2022.

Idea

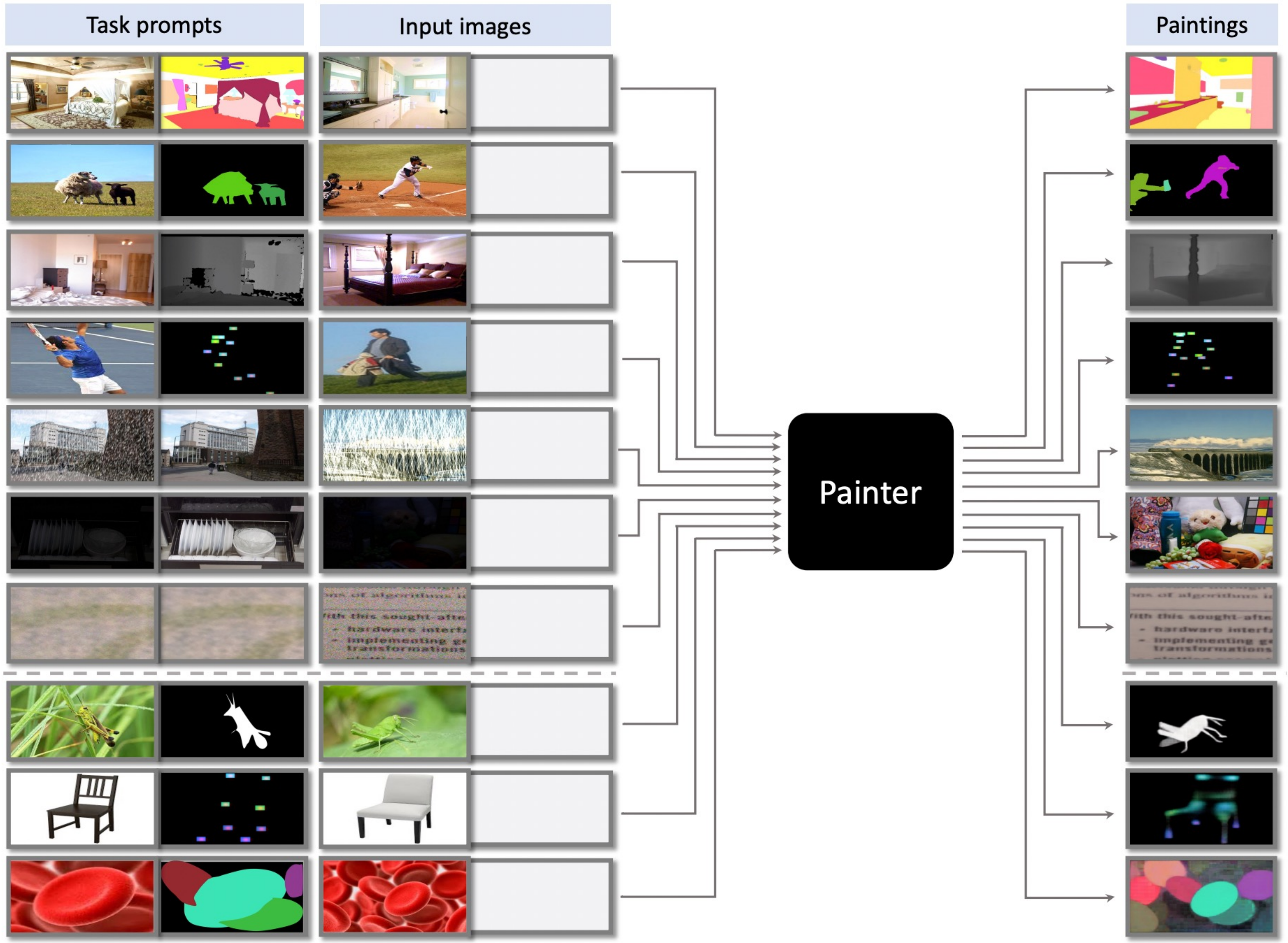
Image inpainting for prompting introduced in 2022 by Bar et al.

Edge detection Colorization Inpainting Segmentation Style transfer

Bar et al. Visual Prompting via Image Inpainting. Neurips 2022.

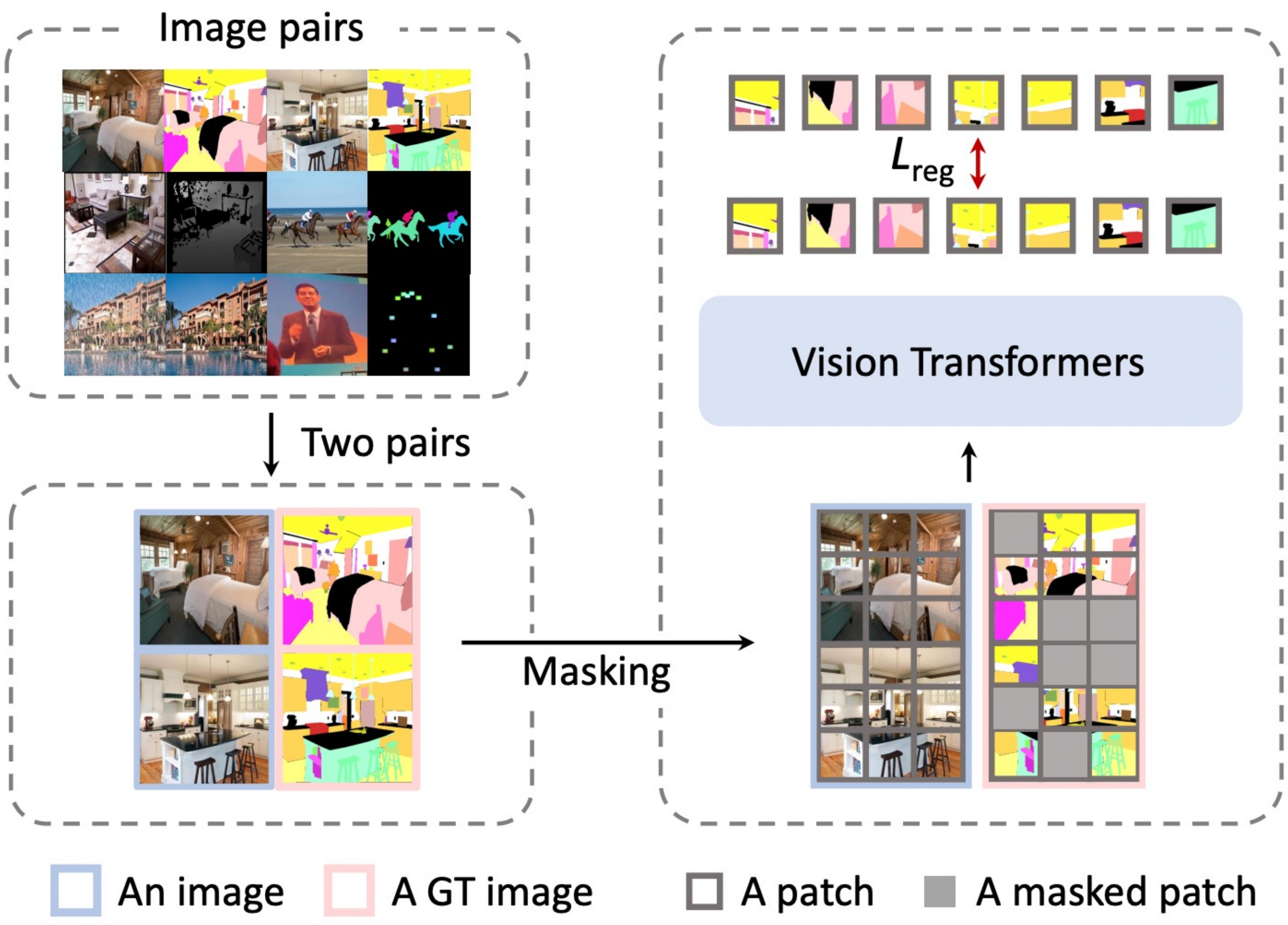
Idea

Idea extended in 2023 by Wang el. on standard vision benchmark datasets



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

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

Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

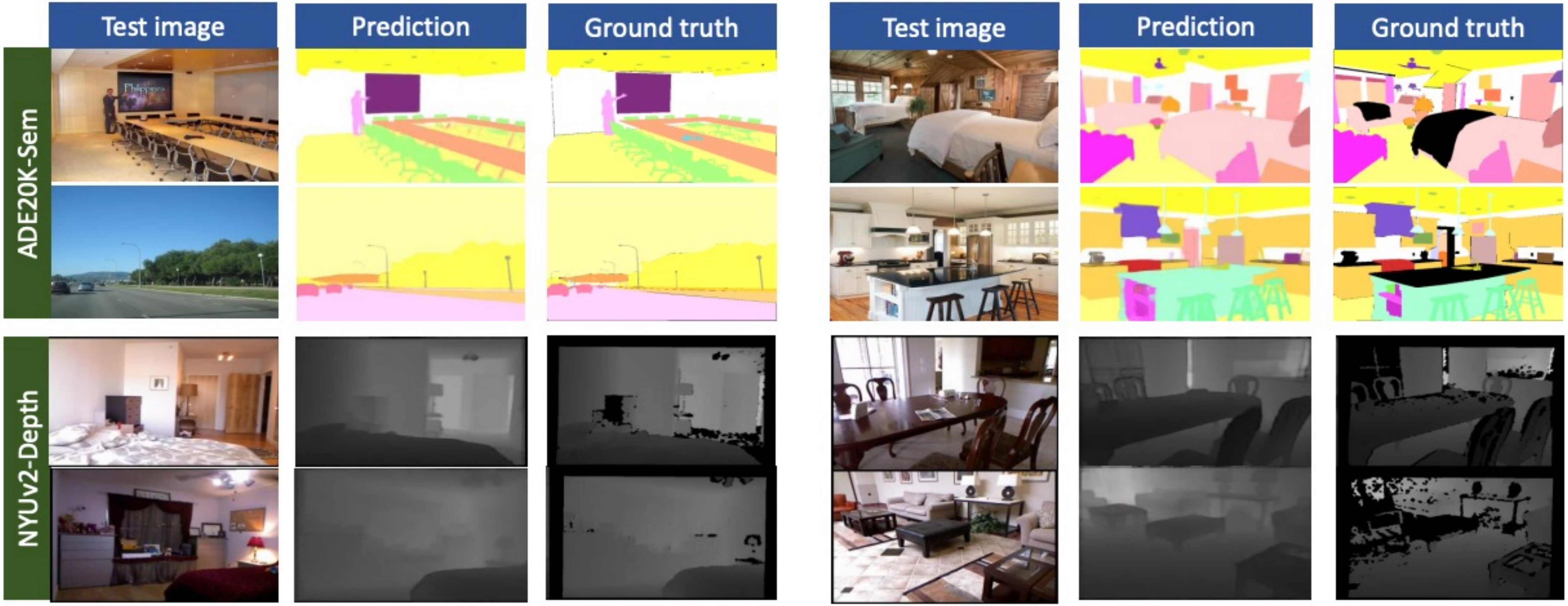
Experimental Results

(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.

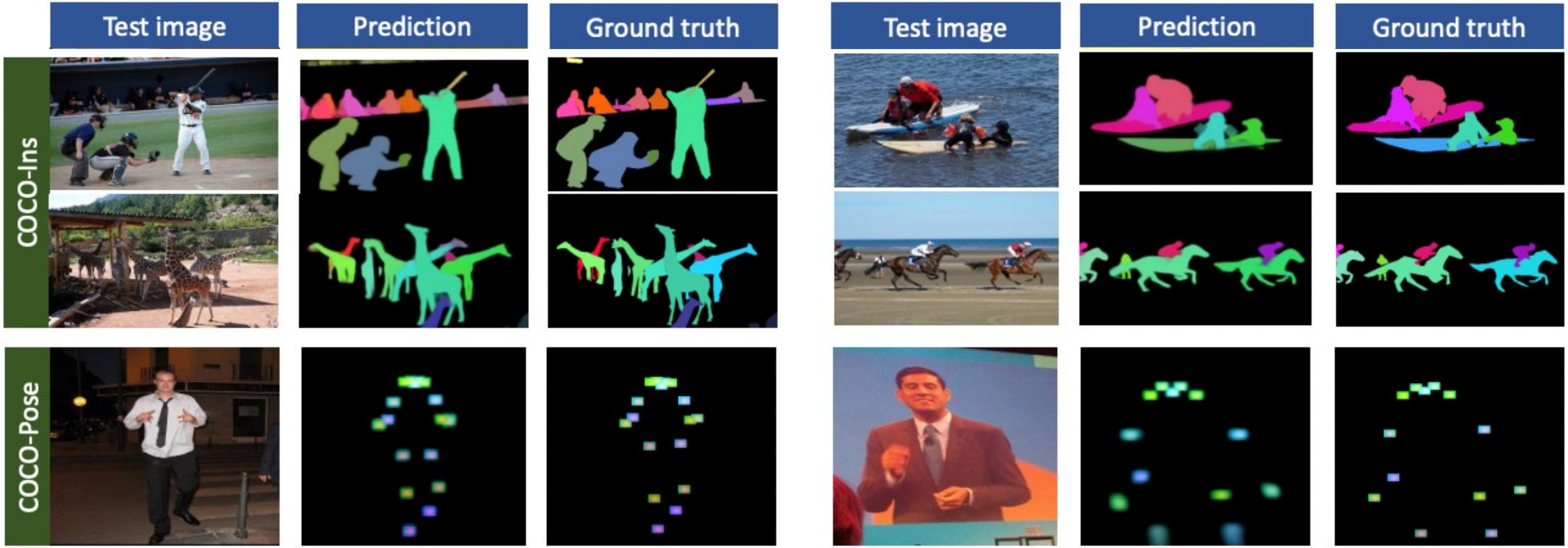
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Qualitative Results: In-Domain Results



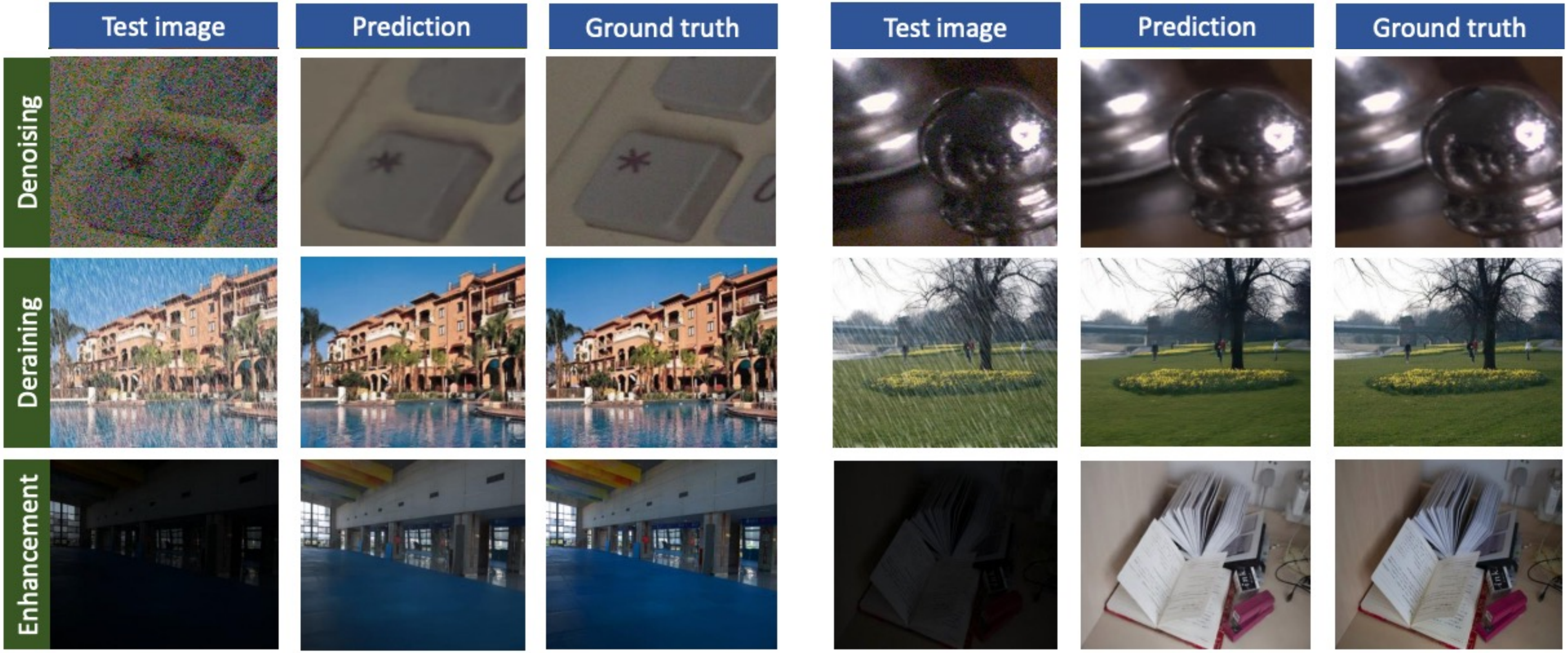
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Qualitative Results: In-Domain Results



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

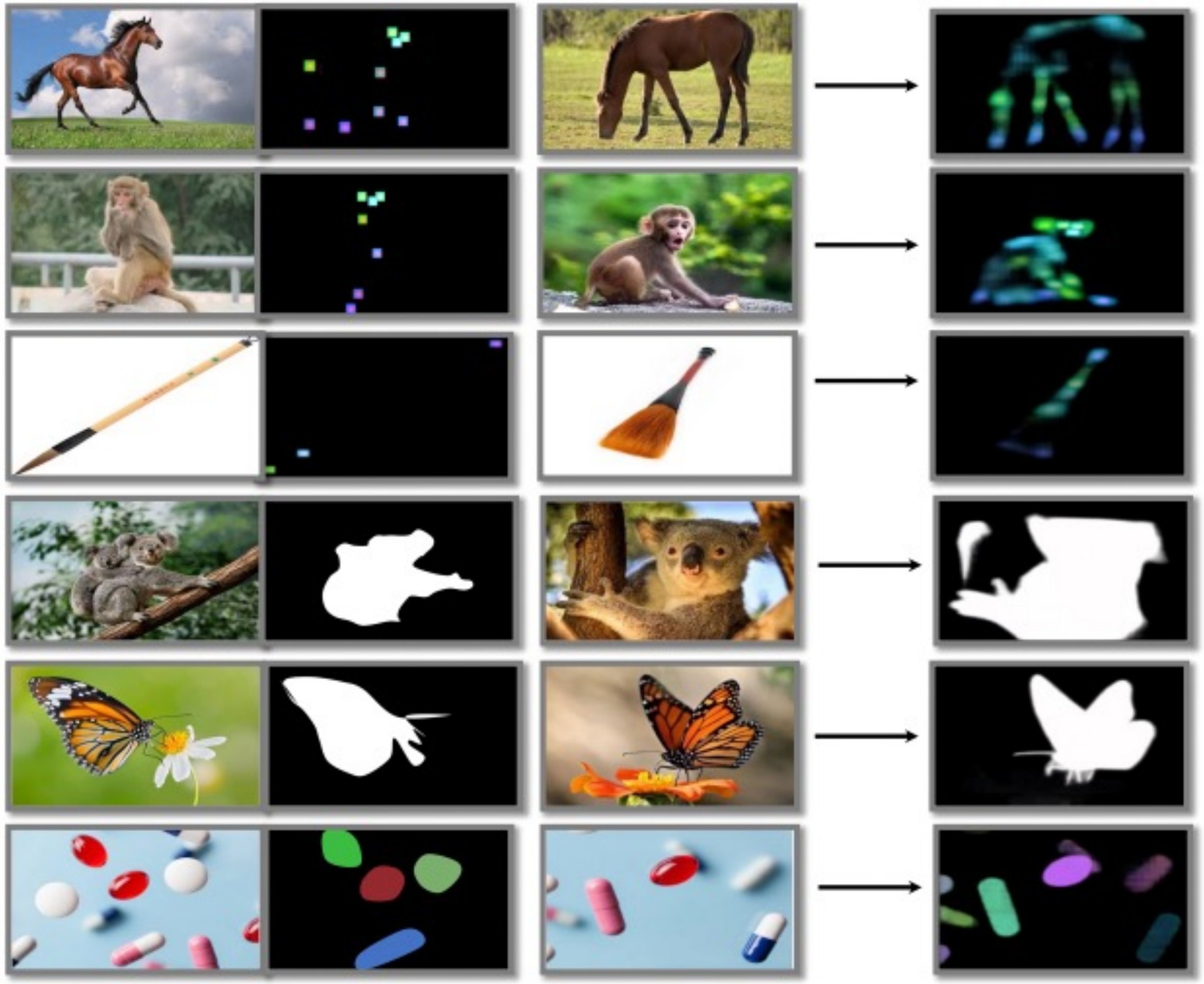
Qualitative Results: In-Domain Results



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

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

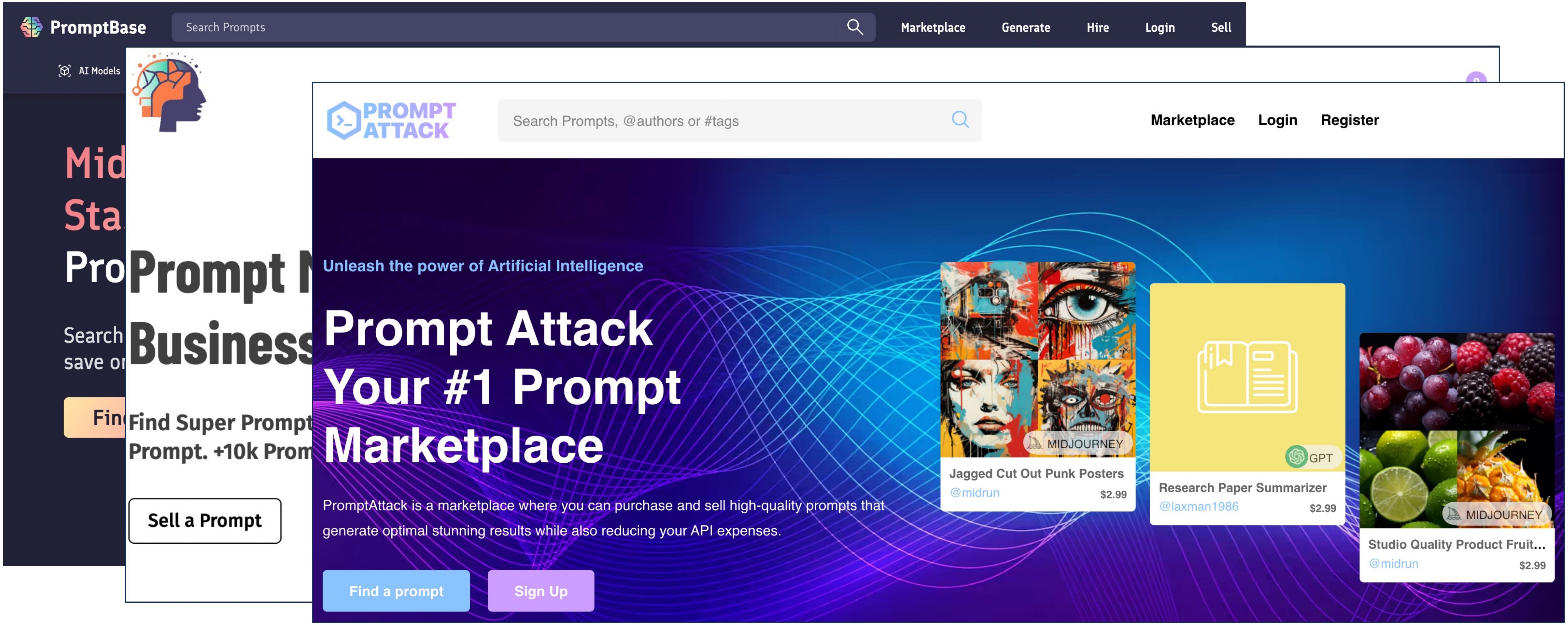


Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Today’s Topics

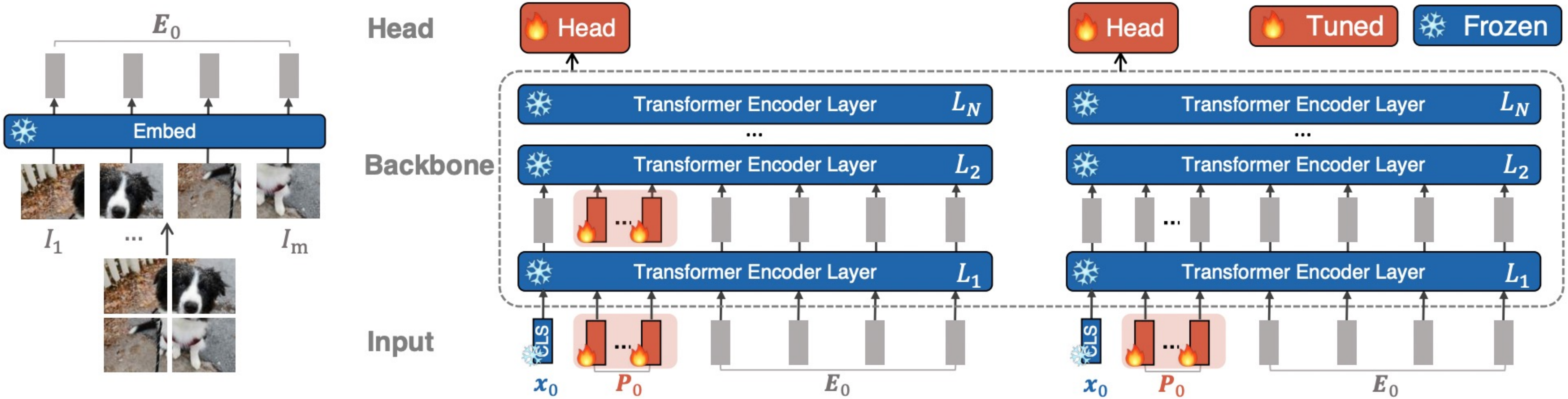
- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

Motivation: Improve Model Performance Without Humans



Manually engineering prompts is challenging to do well (leading to MANY prompt marketplaces)

Idea: Replace Manually-Authored Prompts with Learnable Parameters



(a) Visual-Prompt Tuning: Deep (b) Visual-Prompt Tuning: Shallow

Learned prompts adapt frozen model (e.g., no fine-tuning required) to different target tasks

Jia et al. Visual Prompt Tuning. ECCV 2022.

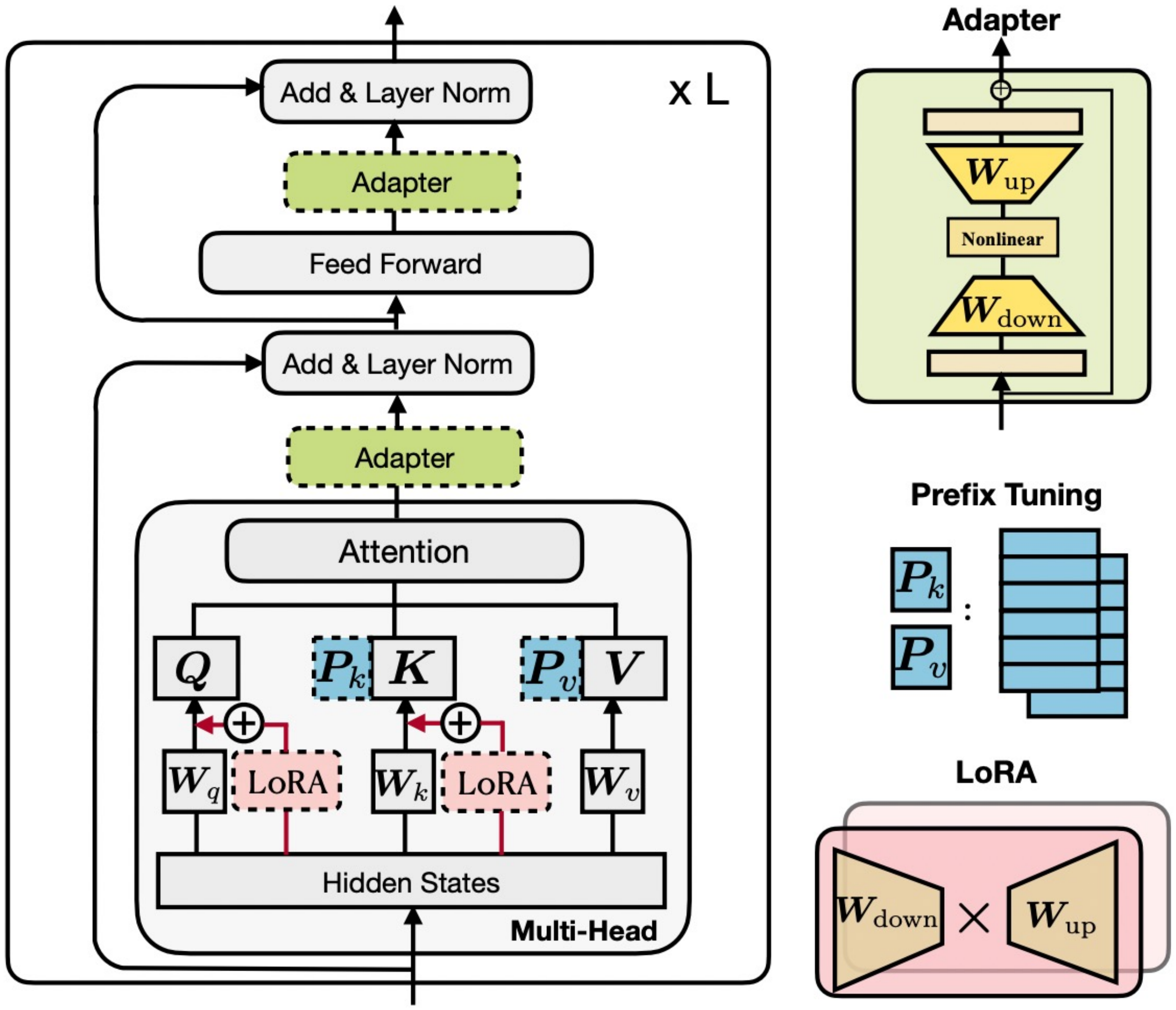


Figure 1: Illustration of the transformer architecture and several state-of-the-art parameter-efficient tuning methods. We use blocks with dashed borderlines to represent the added modules by those methods.

More Generally, Fix Model and Modify Some Added Parameters

(As Shown, There Are Many Approaches)

He et al. Towards a Unified View of Parameter-Efficient Transfer Learning. ICLR 2022.

Competitive Performance With Full Fine-Tuning While Training 15% or Less Parameters

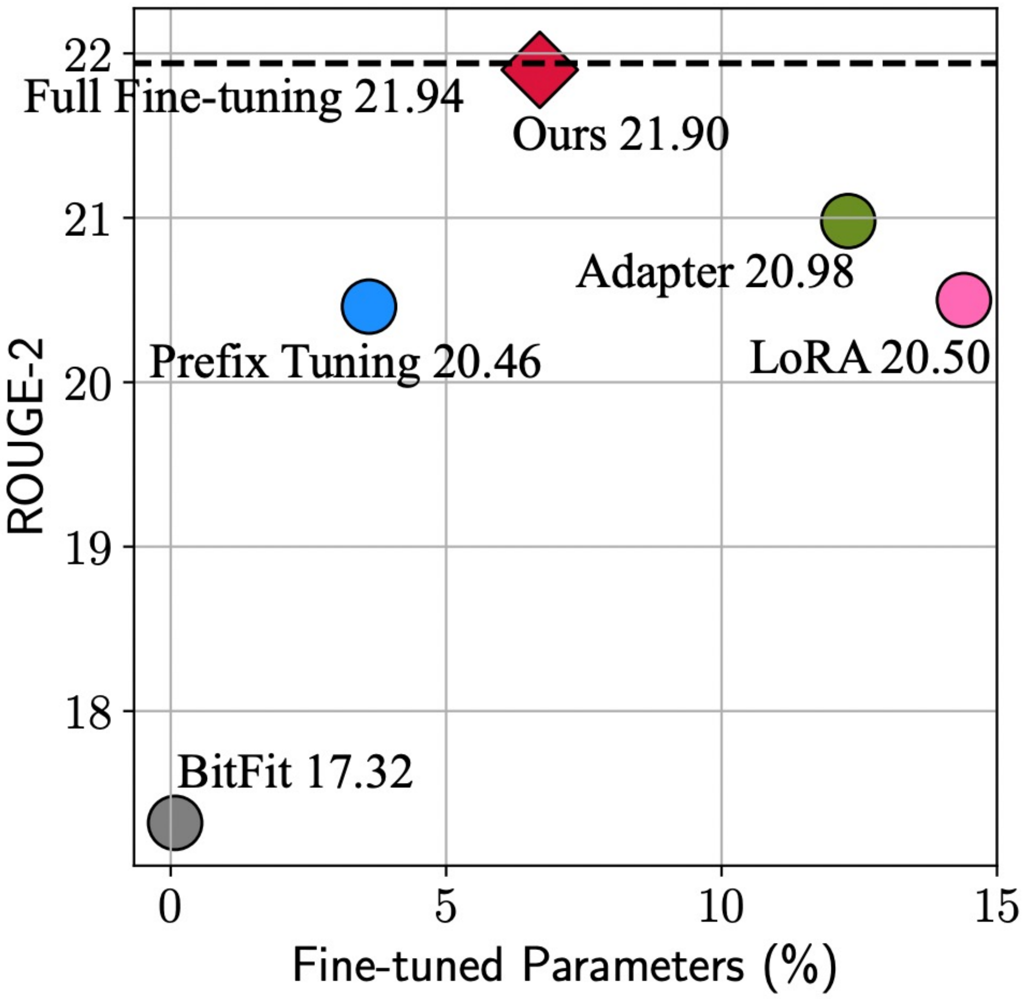


Figure 2: Performance of different methods on the XSum (Narayan et al., 2018) summarization task. The number of fine-tuned parameters is relative to the tuned parameters in full fine-tuning.

He et al. Towards a Unified View of Parameter-Efficient Transfer Learning. ICLR 2022.

What Are Benefits of Tuning Methods?

- Typically, little training data is needed because only a limited amount of parameters need to be trained
- Few task-specific parameters need to be learned and stored to support a new task, compared to model fine-tuning
- Prevents overfitting generalizable knowledge and overfitting to the task
- Provides a static knowledge-base

Jia et al. Visual Prompt Tuning. ECCV 2022.

Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial

Today’s Topics

- Motivation
- Foundation Models
- NLP: Textual Prompting
- CV: Visual Prompting (In-context Few-shot Learning)
- Parameter-Efficient Tuning Methods
- Latex Tutorial


The End