



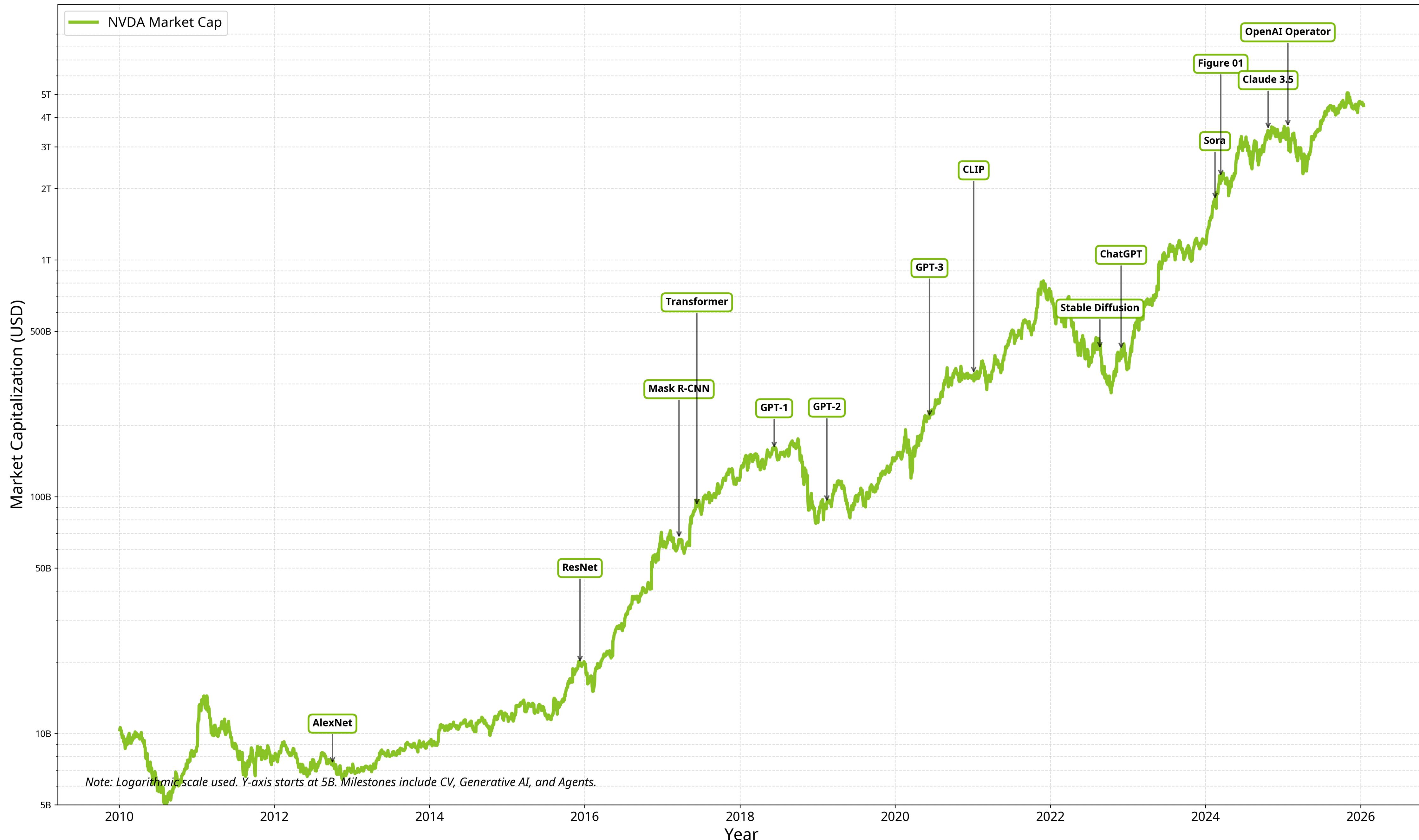
# Multimodal Foundation Models

## Reasoning, Grounding, Awareness, and Adaptation

Kaiyang Zhou  
HKBU CS

**AI has advanced significantly over the past decade**

# Nvidia (NVDA) Market Capitalization & AI Revolution Timeline (2010-2026)



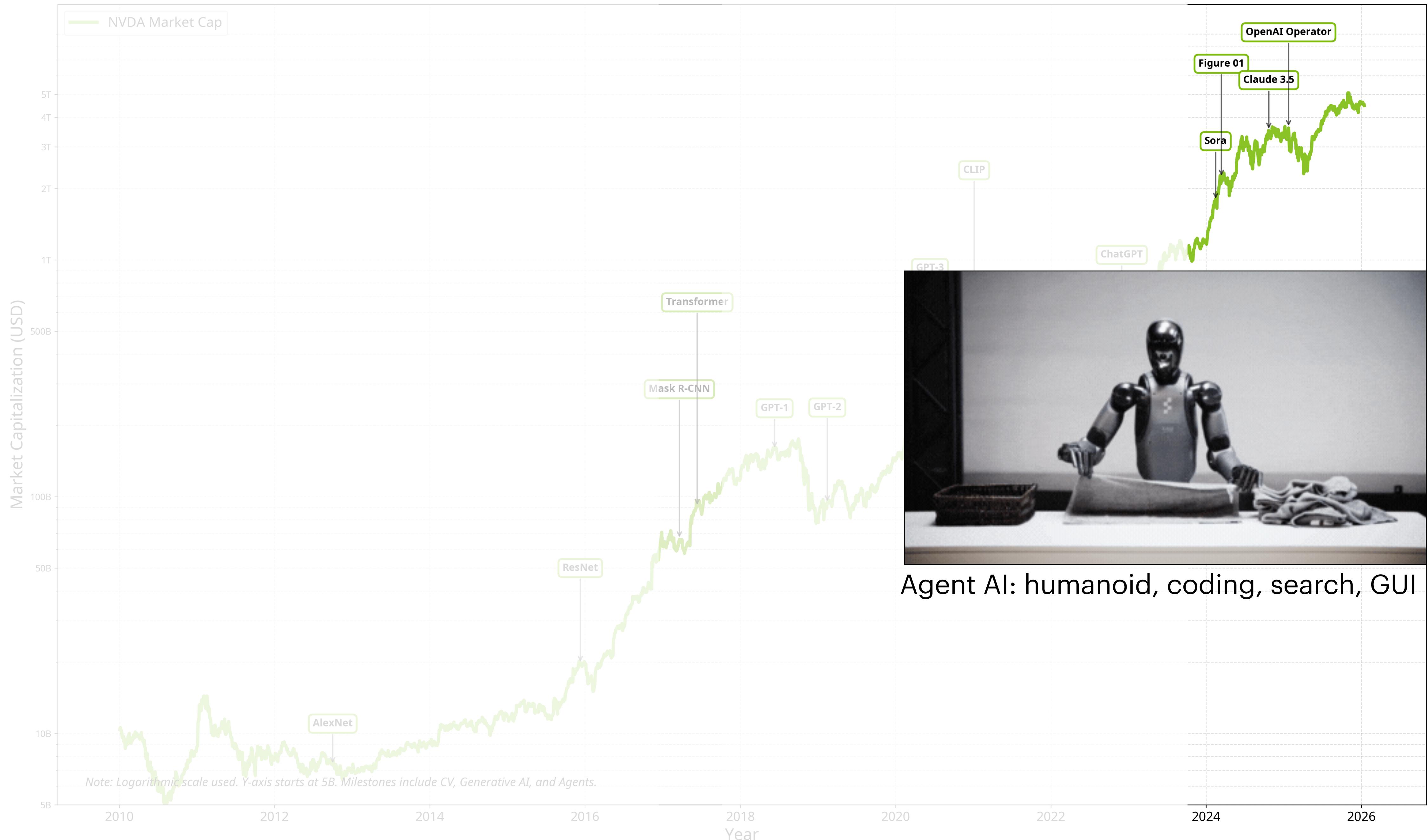
# Nvidia (NVDA) Market Capitalization & AI Revolution Timeline (2010-2026)



# Nvidia (NVDA) Market Capitalization & AI Revolution Timeline (2010-2026)



# Nvidia (NVDA) Market Capitalization & AI Revolution Timeline (2010-2026)



# **Multimodal Foundation Models**

Large Model

Large Vision Model

Large Language Model

Pre-trained Model

Self-Supervised Model

Large Model → focuses too much on size

Large Vision Model      ] → too narrow, only related to vision/language  
Large Language Model      ]

Pre-trained Model → seems like most important things happen after “pre-training”

Self-Supervised Model → too specific to the training objective

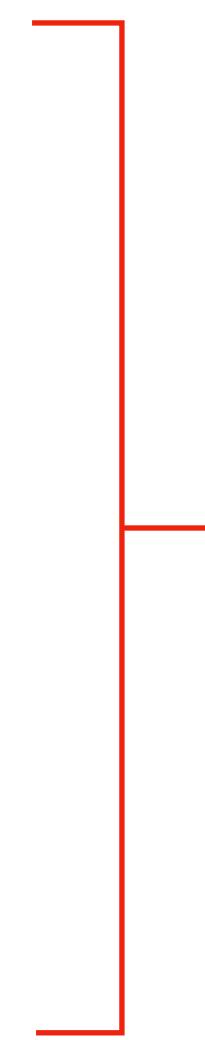
Large Model

Large Vision Model

Large Language Model

Pre-trained Model

Self-Supervised Model



Massive scale (large #params, broad data)

Generality and versatility

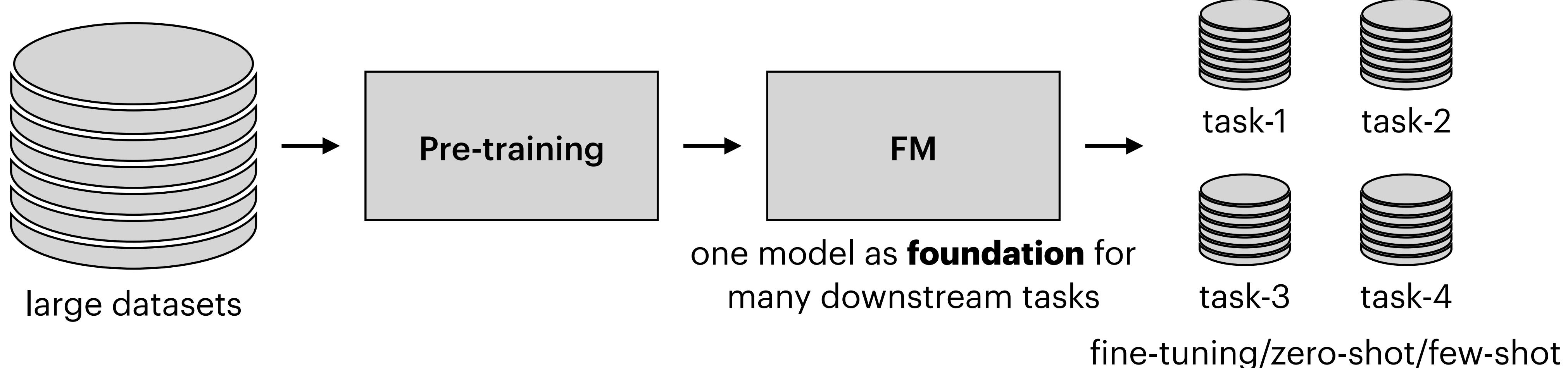
Self-supervised pre-training

Emergent capabilities

# Foundation model

- Trained on broad data, robust to many tasks
- Large #params, large training data, SSL, ...
- “...critically central yet incomplete character”

(Bommasani et al., 2021)



## On the Opportunities and Risks of Foundation Models

Rishi Bommasani\* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora  
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill  
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji  
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue  
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh  
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman  
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt  
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain  
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani  
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi  
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent  
Xiang Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning  
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan  
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan  
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech  
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren  
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh  
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin  
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu  
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia  
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou  
Percy Liang<sup>\*1</sup>

Center for Research on Foundation Models (CRFM)  
Stanford Institute for Human-Centered Artificial Intelligence (HAI)  
Stanford University

# GPT

## Generative Pre-trained Transformer

### Improving Language Understanding by Generative Pre-Training

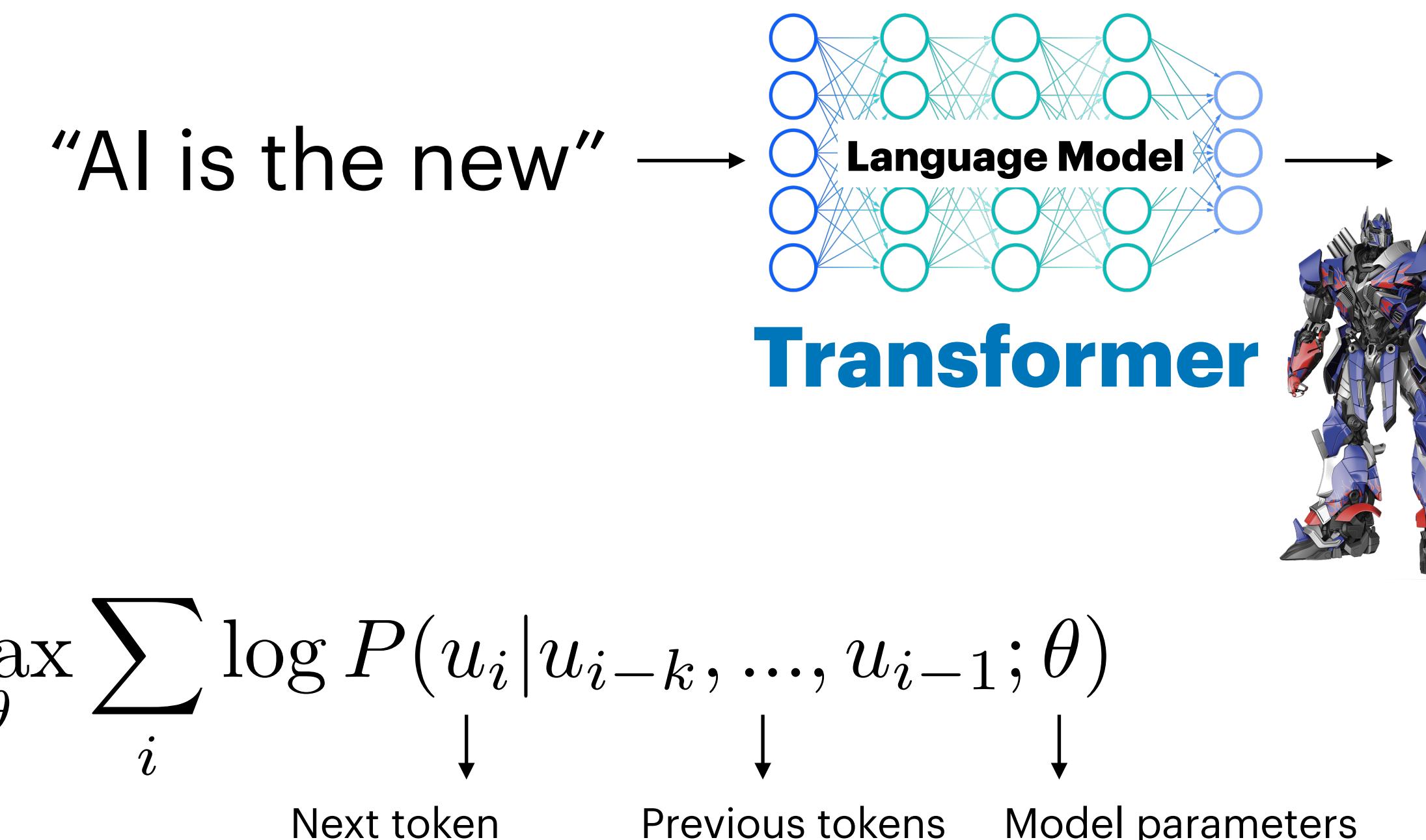
Alec Radford  
OpenAI  
alec@openai.com

Karthik Narasimhan  
OpenAI  
karthikn@openai.com

Tim Salimans  
OpenAI  
tim@openai.com

Ilya Sutskever  
OpenAI  
ilyasu@openai.com

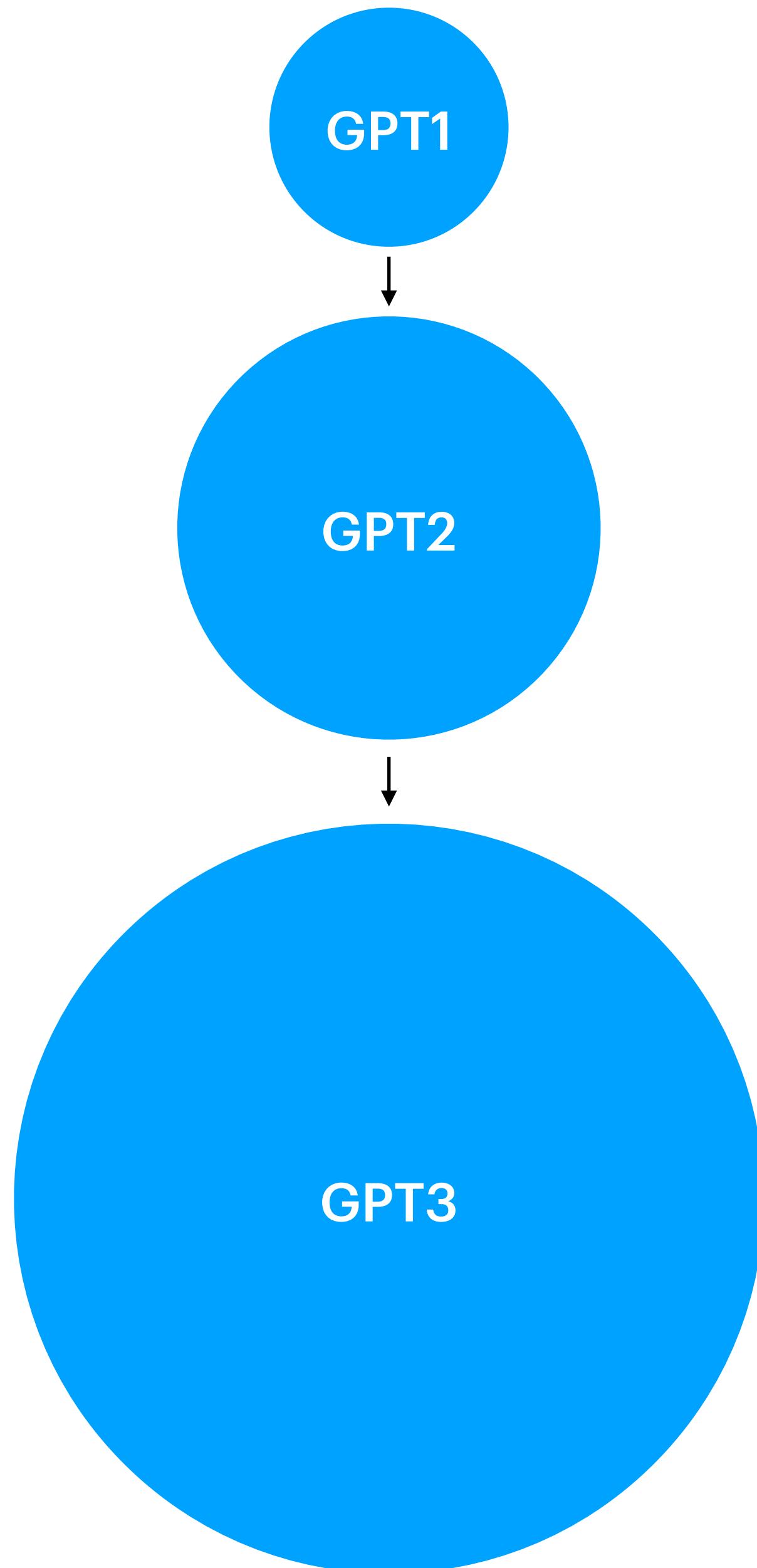
## Autoregressive training (scaled up)



Word	Probability
a	0.000001
ah	0.000002
...	...
elect	0.000022
electricity	0.03
...	...
zip	0.000034

# GPT

- Learning objective: **next-token prediction**.
- GPT1:
  - 12 decoder-only transformer layers, **117M params**
  - 512 context tokens, 768-d hidden size
  - trained on BooksCorpus (>7,000 books)
- GPT2:
  - 48 layers, **1.5B params**
  - 1024 context tokens, 1600-d hidden size
  - trained on larger datasets
- GPT3:
  - 96 layers, **175B params**
  - 2048 context tokens, 12288-d hidden size
  - trained on even larger datasets (300B tokens)



# ...increasingly convincing generations (by GPT2)

Conditional generation on an out-of-distribution context

**Context (human-written):** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

**GPT-2:** The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

# ...increasingly convincing generations (by GPT3)

---

Context →

The City  
BY C. P. CAVAFY  
TRANSLATED BY EDMUND KEELEY  
[Poem text omitted]

SOME TREES  
John Ashbery  
[Poem text omitted]

Shadows on the Way  
Wallace Stevens American modernist poet (1879-1955)

---

----- Generated Poem 1 -----

I must have shadows on the way  
If I am to walk I must have  
Each step taken slowly and alone  
To have it ready made

And I must think in lines of grey  
To have dim thoughts to be my guide  
Must look on blue and green  
And never let my eye forget  
That color is my friend  
And purple must surround me too

----- Generated Poem 3 -----

The sun was all we had. Now, in the shade  
All is changed. The mind must dwell on those  
White fields, that to its eyes were always old;  
Those ancient gleams, convoluting  
The way they lay among the huge roots,  
The great dark tomes of reverie,  
The plumed labyrinth of the sea.

When our own breath gonme faint and fade  
Up the long stairway of the dead  
We shall remember well

# LM ≠ Assistant

Prompt *Explain the moon landing to a 6 year old in a few sentences.*

Completion GPT-3

*Explain the theory of gravity to a 6 year old.*

*Explain the theory of relativity to a 6 year old in a few sentences.*

*Explain the big bang theory to a 6 year old.*

*Explain evolution to a 6 year old.*

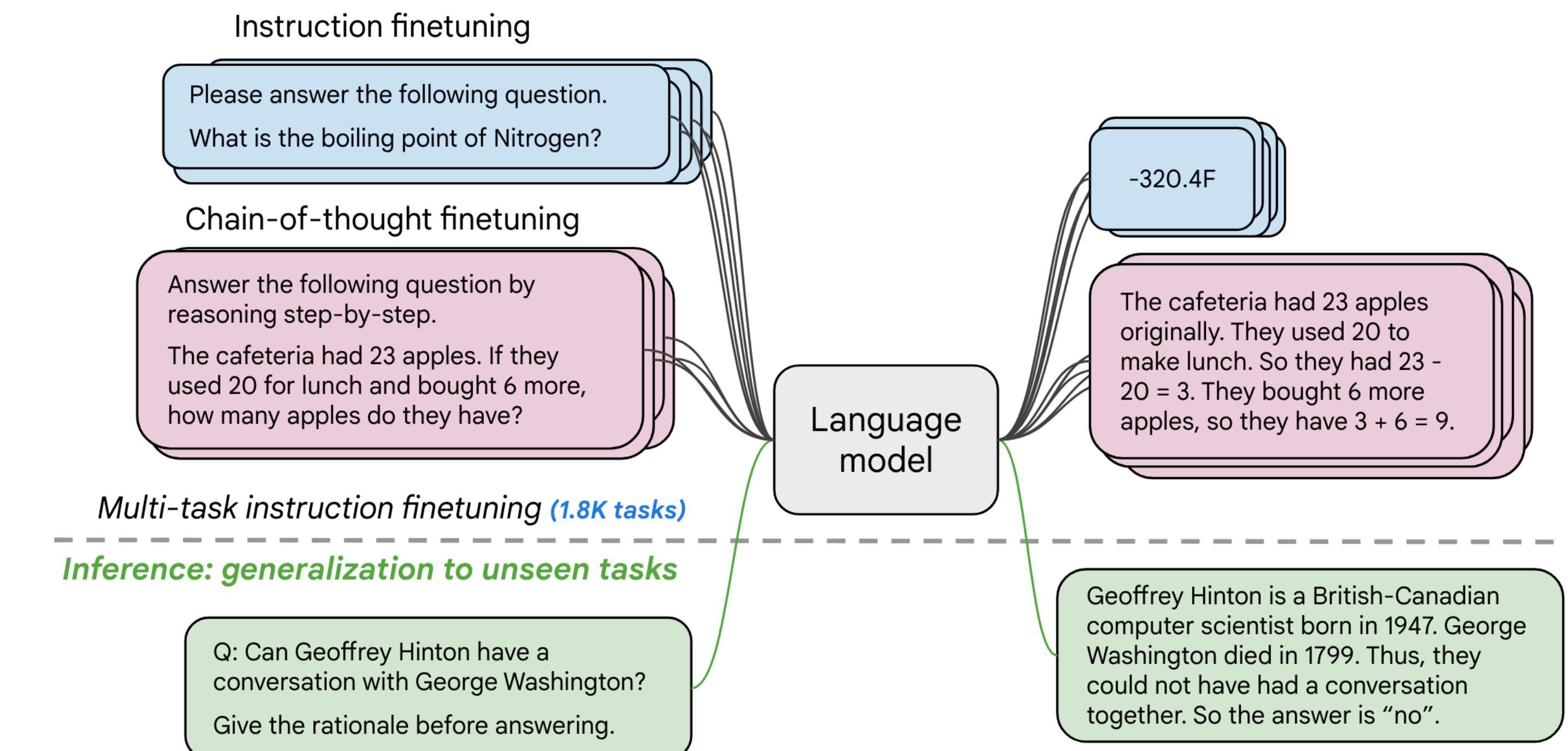
InstructGPT

*People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.*

# Instruction tuning

- = Supervised FT on **instruction data** ( $x, y$ )
  - $x$ : prompt
  - $y$ : completion
- Next-token prediction

$$\text{loss: } - \sum_{i=1}^{|y|} \log P(y_i | y_{<i}, x)$$



# Instruction tuning

- (Optional) Calculate loss on the **output tokens only**, or the **entire input+output**

## 2.2 Instruction Tuning

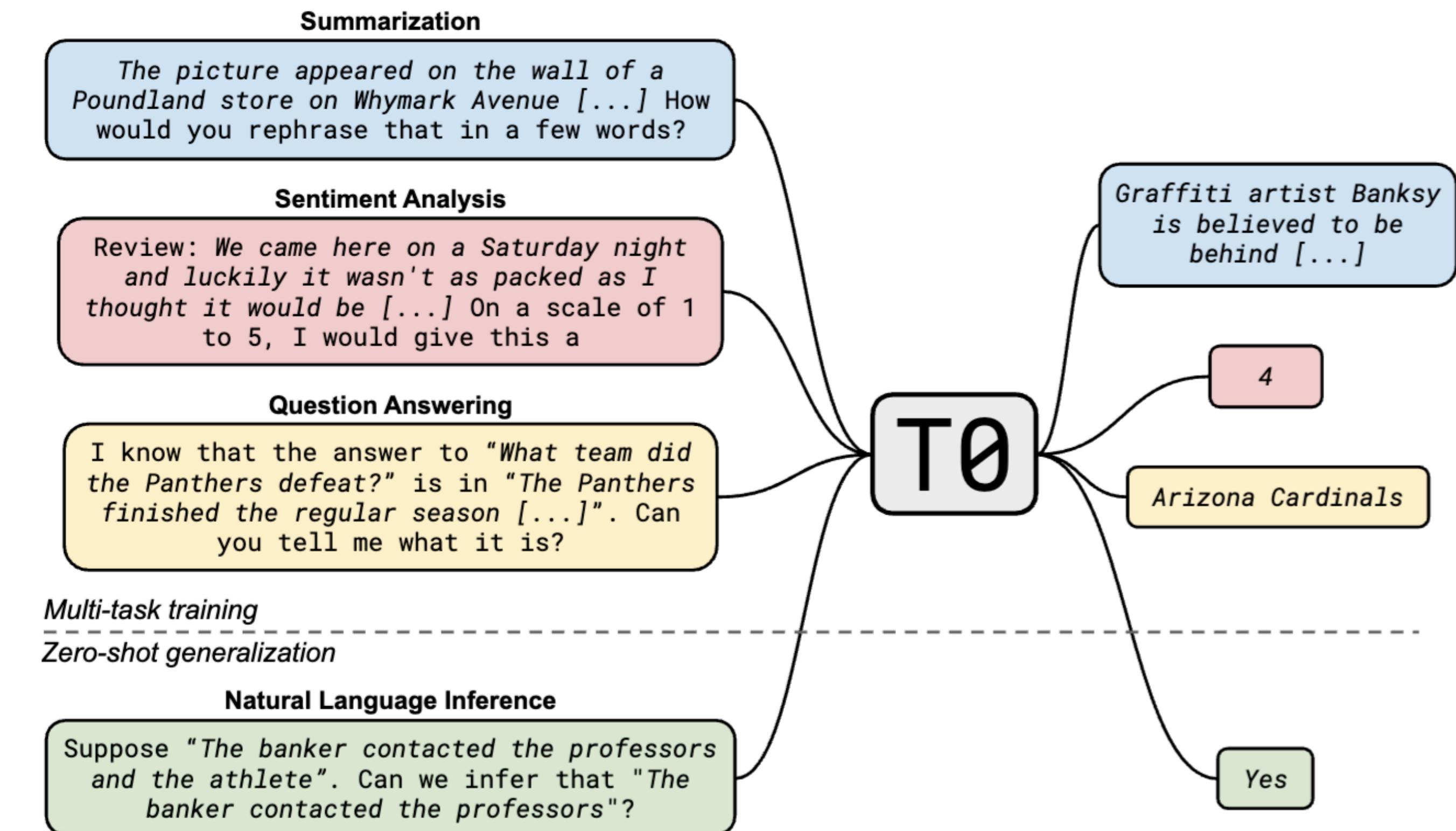
Collecting human demonstration and preference labels for LLM alignment is a cumbersome and expensive process ([Ouyang et al., 2022](#); [Touvron et al., 2023](#)). The challenge and cost are more pronounced under long-context scenarios, which often involve complex information flow and specialized knowledge, e.g., processing dense legal/scientific documents, making the annotation task nontrivial even for skilled annotators. In fact, most existing open-source instruction datasets ([Conover et al., 2023](#); [Köpf et al., 2023](#)) predominantly consist of short samples.

In this work, we found that a simple and cheap approach which leverages a pre-built large and diverse short-prompt dataset works surprisingly well on long-context benchmarks. Specifically, we take the RLHF dataset used in LLAMA 2 CHAT and augment it with synthetic self-instruct ([Wang et al., 2022](#)) long data generated by LLAMA 2 CHAT itself, in the hope that the model can learn a diverse set of skills through the large amount of RLHF data and transfer that knowledge to long-context scenarios via self-instruct data. The data generation process focuses on QA-format tasks: starting from a long document in our pretraining corpus, we select a random chunk and prompt LLAMA 2 CHAT to write question-answer pairs based on information in the text chunk. We collect both long and short form answers with different prompts. After that, we also adopt a self-critique step where we prompt LLAMA 2 CHAT to verify the model-generated answers. Given a generated QA pair, we use the original long document (truncated to fit the model's maximum context length) as the context to construct a training instance.

For short instruction data, we concatenate them as 16,384-token sequences. For long instruction data, we add padding tokens on the right so that models can process each long instance individually without truncation. While standard instruction tuning only calculates loss on the output tokens, we find it particularly beneficial to also calculate the language modeling loss on the long input prompts, which gives consistent improvements on downstream tasks ([Section 4.3](#)).

# Instruction tuning

- Unify NLP tasks
  - train: multi-task learning
  - test: next-token prediction
- Generalizable to unseen tasks

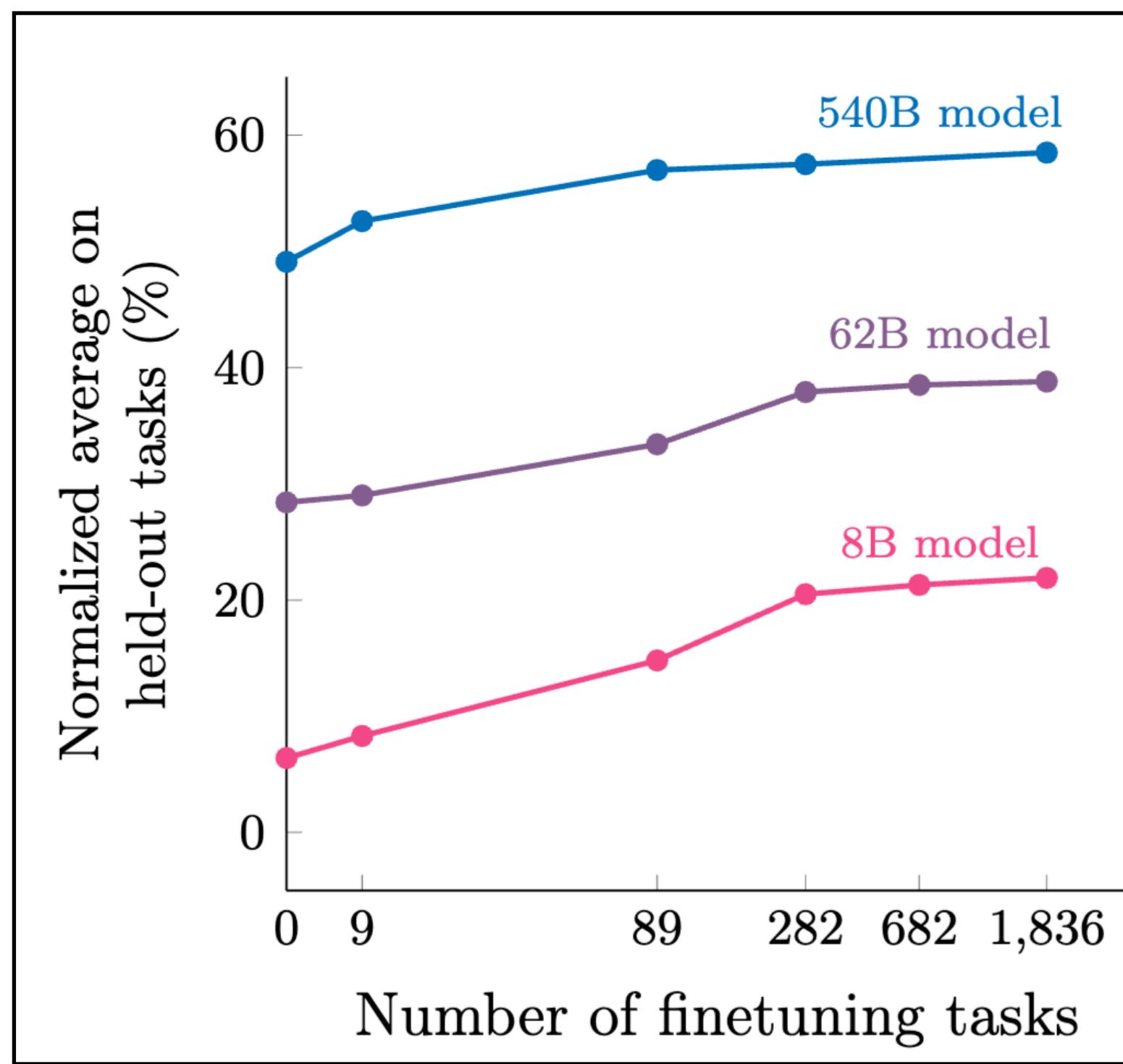


# Instruction tuning

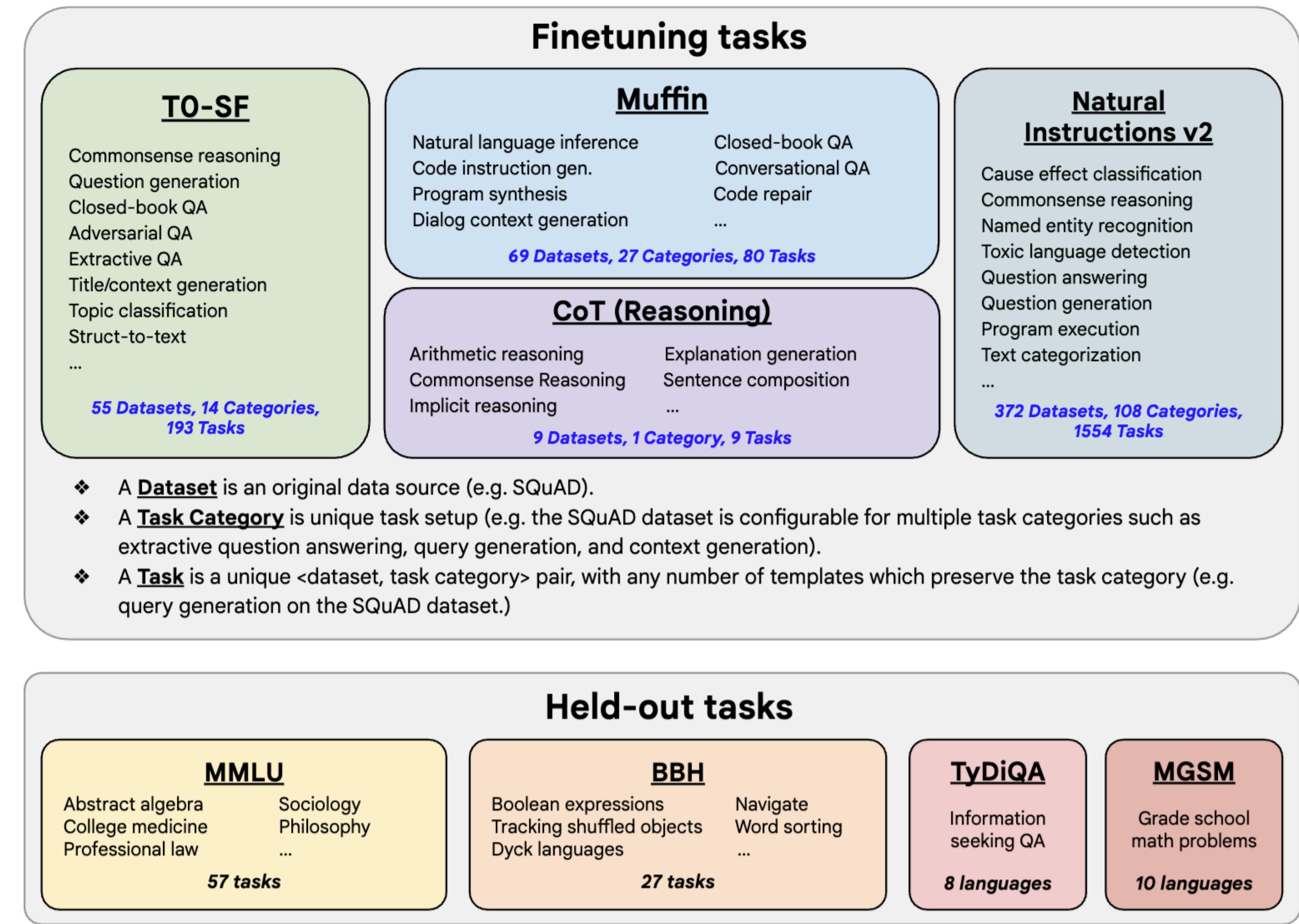
- Data recipe

# Instruction tuning

- Data recipe
  - **scale #tasks**

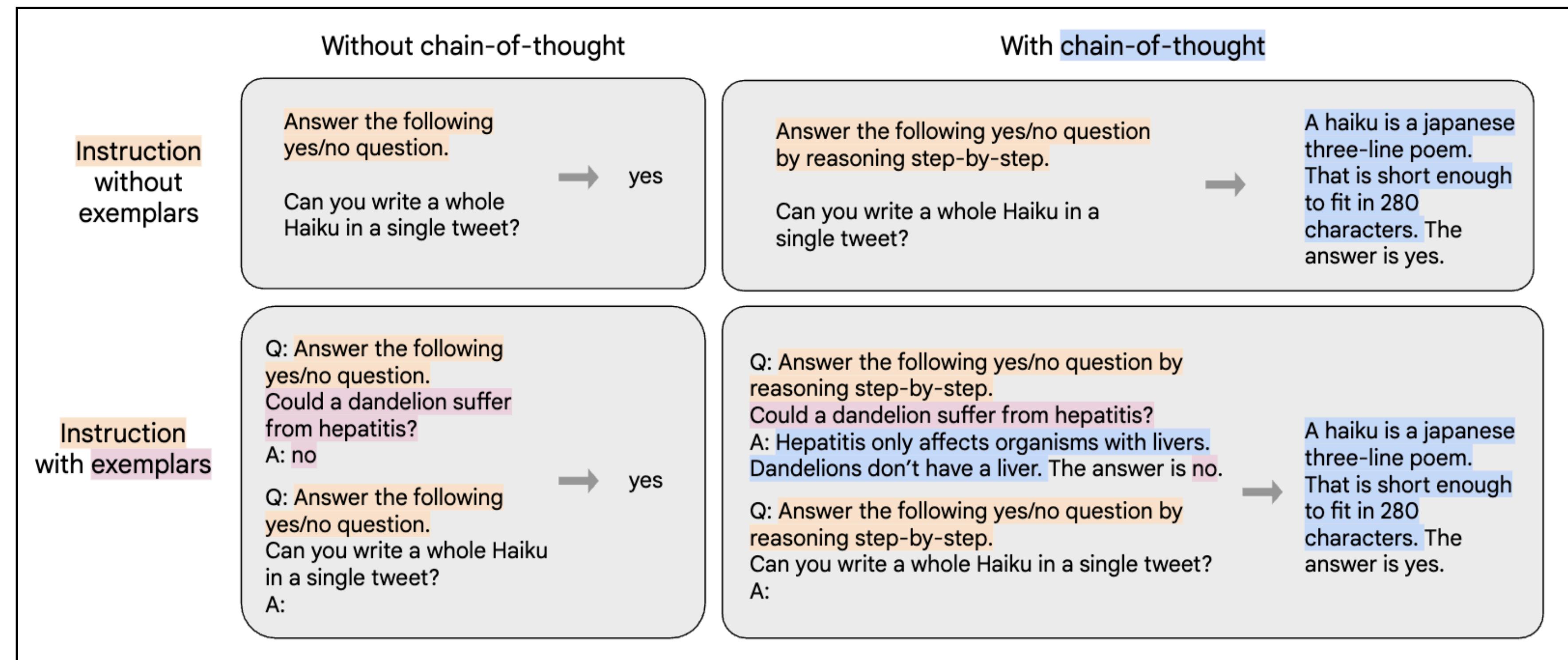


473 datasets, 146 task categories, and 1,836 total tasks



# Instruction tuning

- Data recipe
  - scale #tasks
  - **add CoT & exemplars**



# Instruction tuning

- Data, from where?

# Instruction tuning

- Data, from where?
  - **option #1: human-written from scratch**



“...an open source dataset of instruction-following records generated by thousands of Databricks employees in several of the behavioral categories outlined in the InstructGPT paper, including brainstorming, classification, closed QA, generation, information extraction, open QA, and summarization.”

Datasets: databricks/databricks-dolly-15k like 876 Follow Databricks 1.46k

Tasks: Question Answering Summarization Modalities: Text Formats: json Languages: English Size: 10K - 100K ArXiv:

Libraries: Datasets pandas Croissant +1 License: cc-by-sa-3.0

Dataset card Data Studio Files and versions xet Community 18

Dataset Viewer

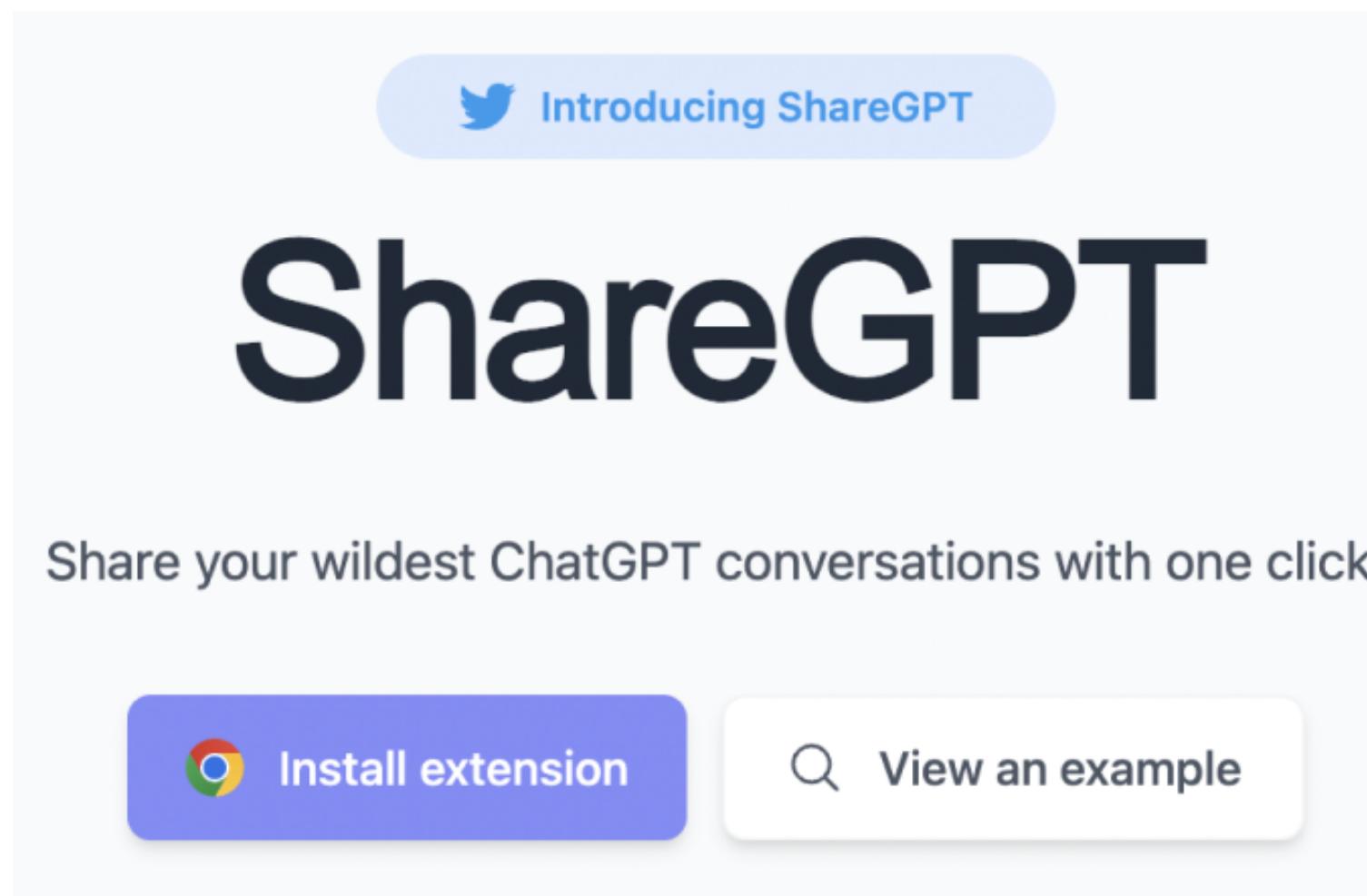
Split (1)  
train · 15k rows

Search this dataset

instruction	context	response	category
string · lengths 4 11.7k	string · lengths 0 23.5k	string · lengths 1 26k	string · classes 8 values
When did Virgin Australia start operating?	Virgin Australia, the trading name of Virgin Australia Airlines Pty...	Virgin Australia commenced services on 31 August 2000 as Virgin Blue,...	closed_qa
Which is a species of fish? Toto or Rope		Toto	classification

# Instruction tuning

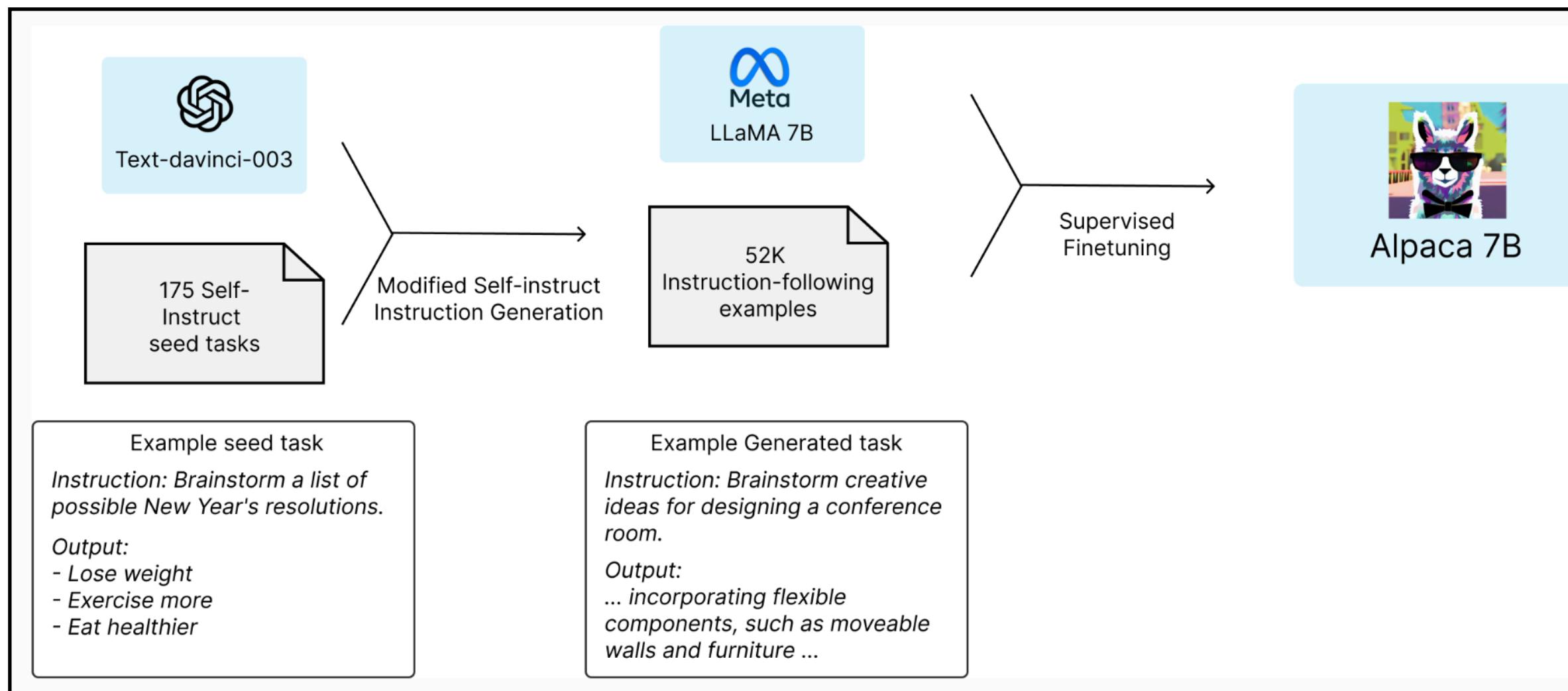
- Data, from where?
  - option #1: human-written from scratch
  - **option #2: human-written prompts + LLM-generated completions (distillation!)**



Vicuna-13B, an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT

# Instruction tuning

- Data, from where?
  - option #1: human-written from scratch
  - option #2: human-written prompts + LLM-generated completions (distillation!)
  - **option #3: all LLM-generated**



"For the data, we generated instruction-following demonstrations by building upon the self-instruct method. We started with the 175 human-written instruction-output pairs from the self-instruct seed set. We then **prompted text-davinci-003 to generate more instructions using the seed set as in-context examples**. We improved over the self-instruct method by simplifying the generation pipeline (see details in GitHub) and significantly reduced the cost. Our data generation process results in 52K unique instructions and the corresponding outputs, which costed less than \$500 using the OpenAI API."  
— from <https://crfm.stanford.edu/2023/03/13/alpaca.html>

# Instruction tuning

- Data quantity: more is better?

---

## LIMA: Less Is More for Alignment

---

Chunting Zhou<sup>μ\*</sup> Pengfei Liu<sup>π\*</sup> Puxin Xu<sup>μ</sup> Srinivas Iyer<sup>μ</sup> Jiao Sun<sup>λ</sup>  
Yuning Mao<sup>μ</sup> Xuezhe Ma<sup>λ</sup> Avia Efrat<sup>τ</sup> Ping Yu<sup>μ</sup> Lili Yu<sup>μ</sup> Susan Zhang<sup>μ</sup>  
Gargi Ghosh<sup>μ</sup> Mike Lewis<sup>μ</sup> Luke Zettlemoyer<sup>μ</sup> Omer Levy<sup>μ</sup>

<sup>μ</sup> Meta AI  
<sup>π</sup> Carnegie Mellon University  
<sup>λ</sup> University of Southern California  
<sup>τ</sup> Tel Aviv University

### Abstract

Large language models are trained in two stages: (1) unsupervised pretraining from raw text, to learn general-purpose representations, and (2) large scale instruction tuning and reinforcement learning, to better align to end tasks and user preferences. We measure the relative importance of these two stages by training LIMA, a 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully curated prompts and responses, without any reinforcement learning or human preference modeling. LIMA demonstrates remarkably strong performance, learning to follow specific response formats from only a handful of examples in the training data, including complex queries that range from planning trip itineraries to speculating about alternate history. Moreover, the model tends to generalize well to unseen tasks that did not appear in the training data. In a controlled human study, responses from LIMA are either equivalent or strictly preferred to GPT-4 in 43% of cases; this statistic is as high as 58% when compared to Bard and 65% versus DaVinci003, which was trained with human feedback. Taken together, these results strongly suggest that almost all knowledge in large language models is learned during pretraining, and only limited instruction tuning data is necessary to teach models to produce high quality output.

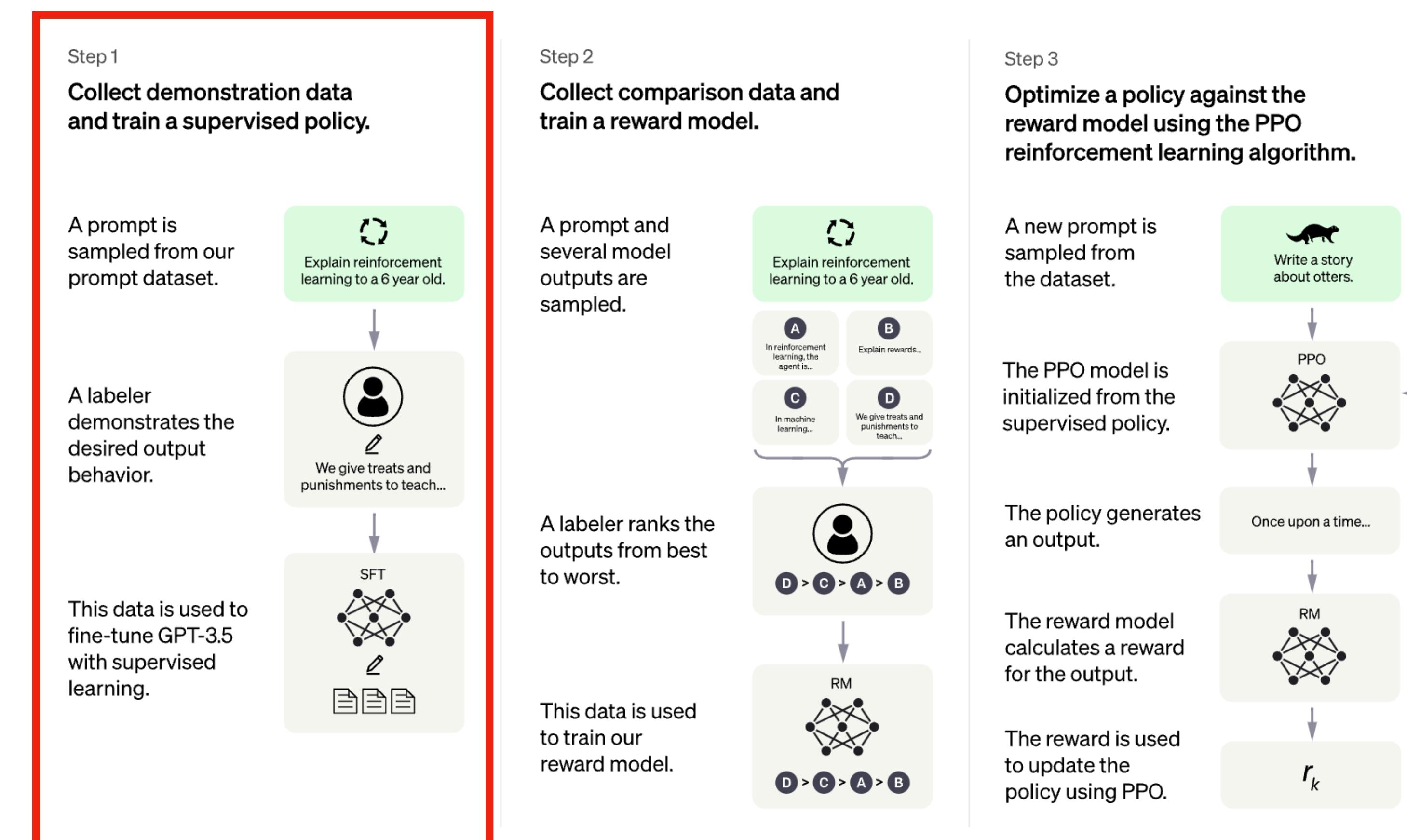
### Key messages:

- Knowledge is learned during pre-training; instruction tuning teaches which sub-distribution of formats to use
- **Quality and diversity** matter: 1,000 manually selected examples work great!

Source	#Examples	Avg Input Len.	Avg Output Len.
<b>Training</b>			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
<b>Dev</b>			
Paper Authors (Group A)	50	36	N/A
<b>Test</b>			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

# Instruction tuning

- Commonly adopted as the first stage of post-training



# **From instruction tuning to alignment**

**Instruction tuning**

**Alignment**

# From instruction tuning to alignment

## Instruction tuning

- execute a specific, given instruction
- often achieved with supervised FT on instruction data

## Alignment

- ensure the model's behaviour aligns with human intent and **values**
- often achieved with RL to align the model with **human preferences**

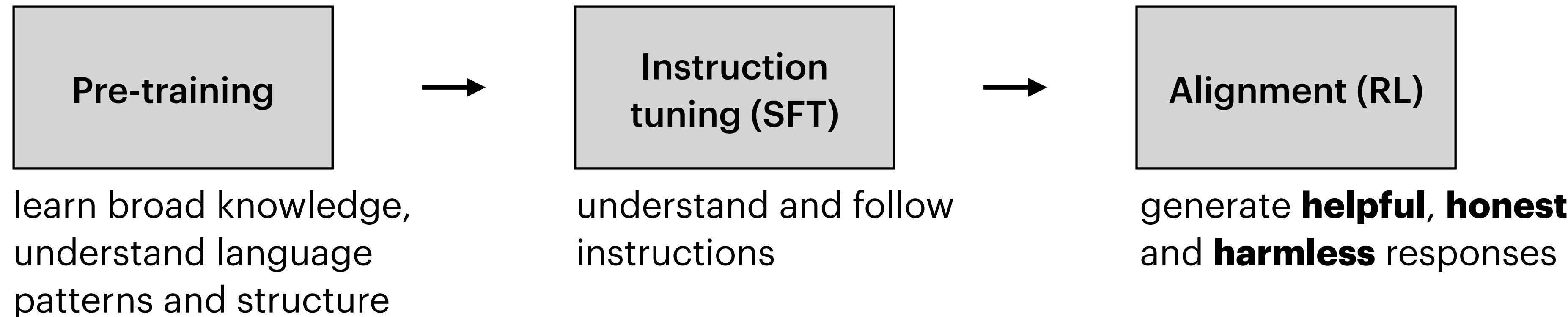
# From instruction tuning to alignment

## Instruction tuning

- execute a specific, given instruction
- often achieved with supervised FT on instruction data

## Alignment

- ensure the model's behaviour aligns with human intent and **values**
- often achieved with RL to align the model with **human preferences**

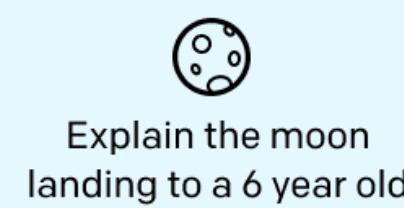


# InstructGPT

Step 1

**Collect demonstration data, and train a supervised policy.**

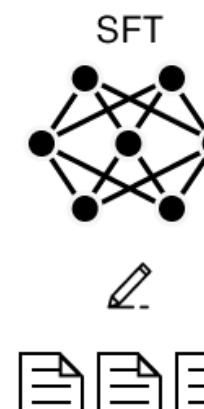
A prompt is sample from our prompt dataset.



A labeler demonstrates the desired output behavior.



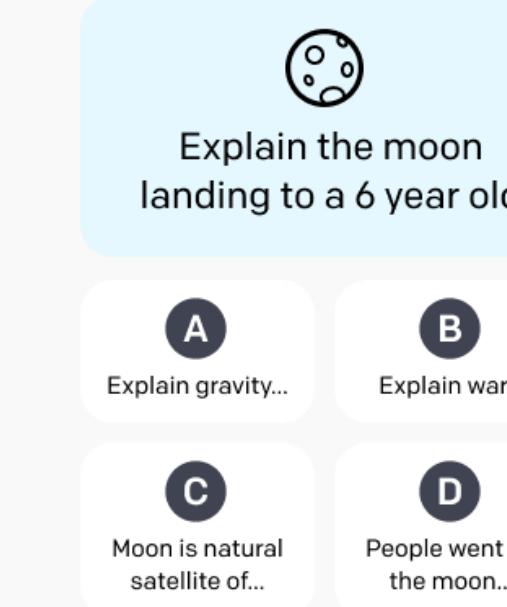
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

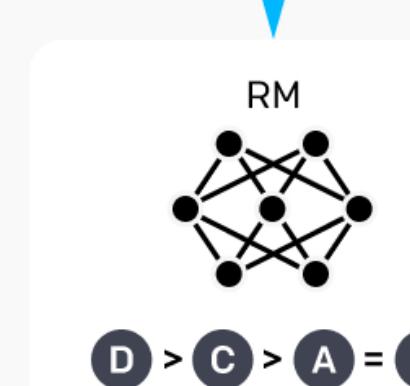
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



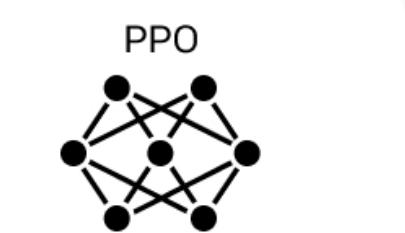
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

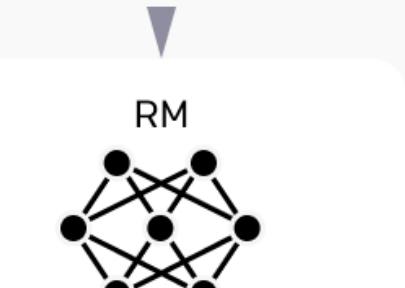
A new prompt is sampled from the dataset.



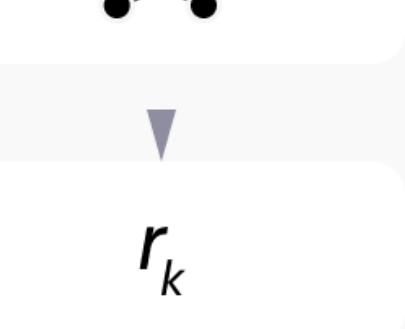
The policy generates an output.



Once upon a time...



The reward model calculates a reward for the output.



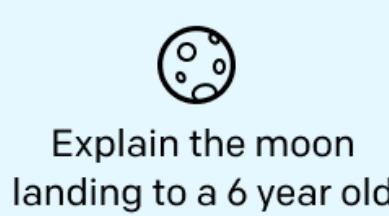
The reward is used to update the policy using PPO.

# InstructGPT

Step 1

Collect demonstration data,  
and train a supervised policy.

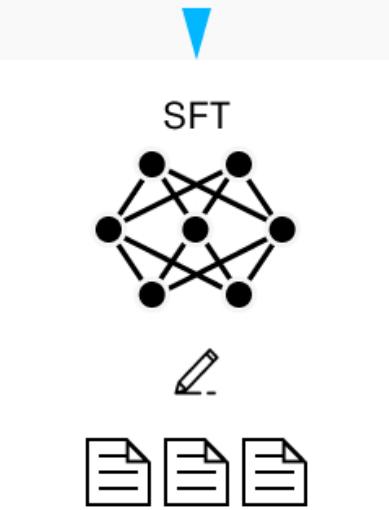
A prompt is sample from  
our prompt dataset.



A labeler demonstrates  
the desired output  
behavior.



This data is used to  
fine-tune GPT-3 with  
supervised learning.



## Supervised FT

- 13k prompts from API and labelers
- Responses written by labelers
- Use cases: generation, open QA, brainstorm, chat, rewrite, summarization, ...

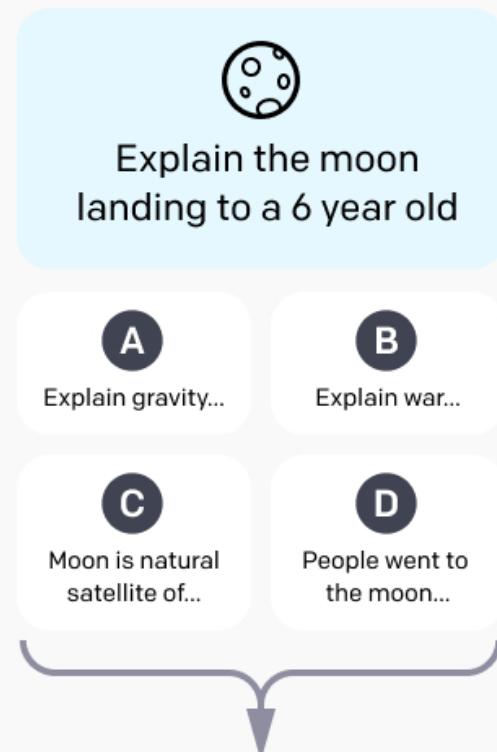
Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

# InstructGPT

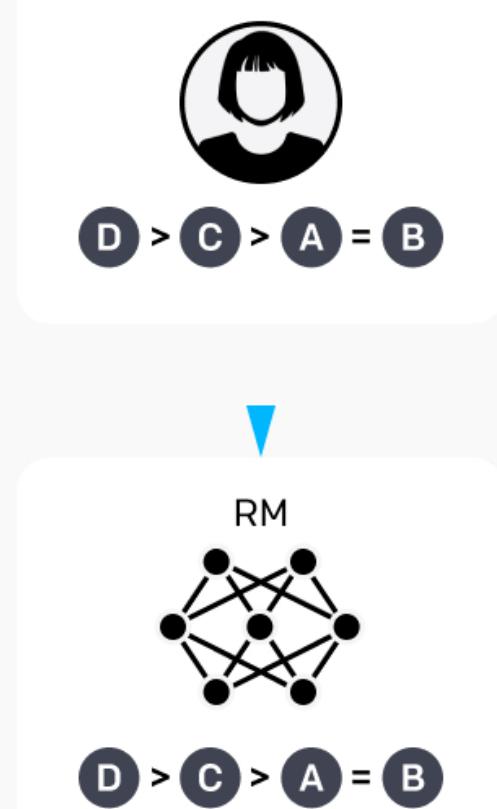
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.

## Reward modeling

- 33k prompts from API and labelers
- Responses generated by the model

“most of our comparison data comes from our supervised policies, with some coming from our PPO policies”

- Labelers rank K responses ( $K=4\sim 9$ )
- 6B reward model:  $(x, y) \rightarrow \mathbb{R}$

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log (\sigma (r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

prompt, winning response, losing response

reward of winning response

reward of losing response

## Ranking outputs

### To be ranked

**B** A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

**C** Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

### Rank 1 (best)

**A** A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

### Rank 2

### Rank 3

**E** Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

### Rank 4

**D** Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

### Rank 5 (worst)

"After evaluating each output individually, labelers rank all the outputs for a given prompt. Ties are encouraged in cases where two outputs seem to be of similar quality"

# InstructGPT

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

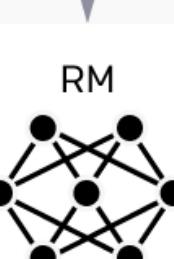


Write a story about frogs



The policy generates an output.

Once upon a time...



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

## RL

- Key idea: FT the supervised policy to maximise rewards (output from RM) using PPO
- 31k prompts from API
- Tip #1: add per-token KL penalty between “PPO” and “SFT” to mitigate over-optimization
- Tip #2: add pre-training loss to “fix the performance regressions on public NLP dataset”

$$\text{objective } (\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} [r_{\theta}(x, y) - \beta \log(\pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x))] + \gamma E_{x \sim D_{\text{pretrain}}} [\log(\pi_{\phi}^{\text{RL}}(x))] + \text{KL penalty}$$

RL reward

a.k.a. reference model

pre-training loss

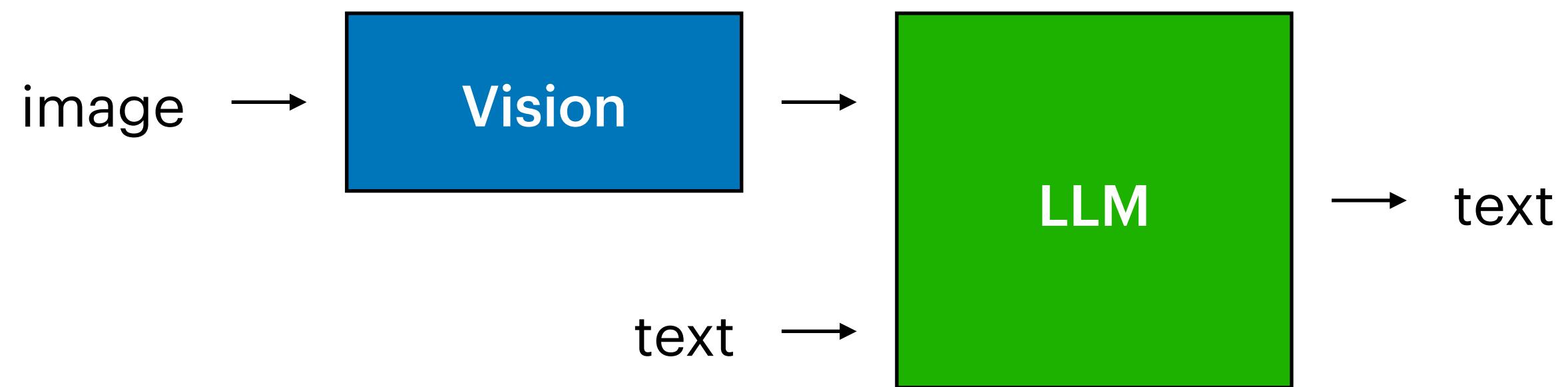
Can see, but can't talk



Can talk, but can't see



Can see, can talk!





natural



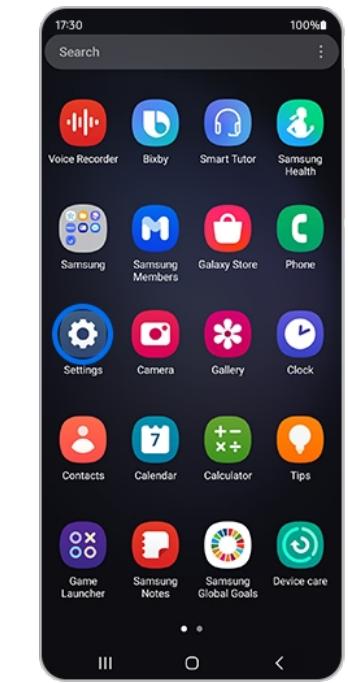
egocentric

Amazon Music [Android] May 25, 2017  
Amazon.com Free ★★★★☆ 12,934

Amazon Drive Feb 27, 2017  
Amazon.com Free ★★★★☆ 1,312

Amazon Alexa Amazon.com Free ★★★★☆ 1,312

GUI (web)



GUI (mobile)

**Q&A: Access to Education**

**Information for field staff and refugee parents**

The following Q&A was produced in accordance with information from the Ministry of Education, Research, and Religious Affairs' plan to ensure better communication with refugee parents and children staying in Greece.

Accessing in public schools in Greece does not obligate refugees to stay in Greece. Education for refugee children is available while they stay in Greece and is beneficial to them, as it provides some stability and normalcy. In addition, documentation of attendance will be provided upon departure from Greece.

**Who is eligible?**

All children have the right to access school education in Greece, without distinction. The unique condition for children to attend school is to be vaccinated, which lies under the responsibility of the Greek Ministry of Health.

The first stage of the Ministry's programme is focusing on providing access to children between 4-15 years old, who are of compulsory school age — kindergarten (σπιaggia) to junior high school (gymnasio).

The Ministry's plan primarily targets the estimated 12,000 refugee and migrant children of compulsory school age (1.4% of the total student population in Greece). In order to accommodate all the refugee children in Greek school systems, the Ministry has identified 1,000 schools by the Ministry. This is for children living in open temporary sites; this record is for children living in urban and semi-urban settings such as relocation accommodation, squats, apartments, hotels, and reception centers for asylum seekers and UAMs.

**Children living in open temporary sites:**

Children between 4-5 years old will be eligible to attend additional kindergarten facilities, which will be established within the open temporary site.

Children aged between 6-15 years will be enrolled in afternoon reception classes from 14:00 to 18:00, in neighbouring public schools identified by the Ministry. They will be taught Greek as a second language, English language, mathematics, sports, arts and computer science.

The International Organization for Migration (IOM) will ensure the transportation of the refugee children from the open temporary sites to the selected schools with buses equipped with IOM escorts.

According to the Greek school system and the new ministerial decision  
According to the Greek school system and the new ministerial decision

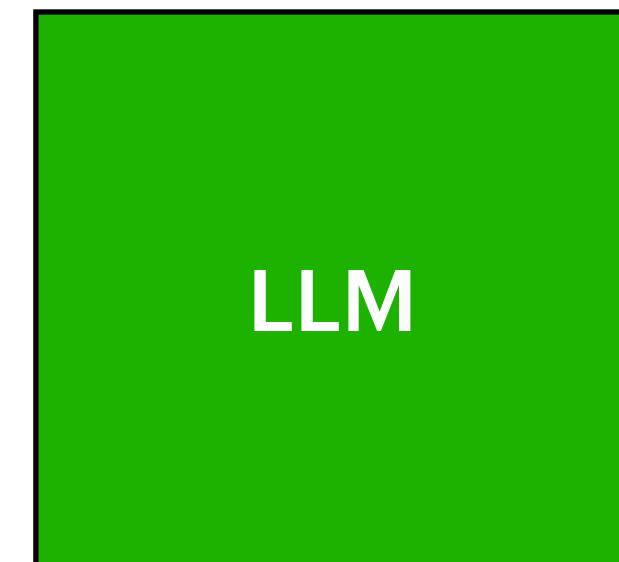
document

Can see, can talk!

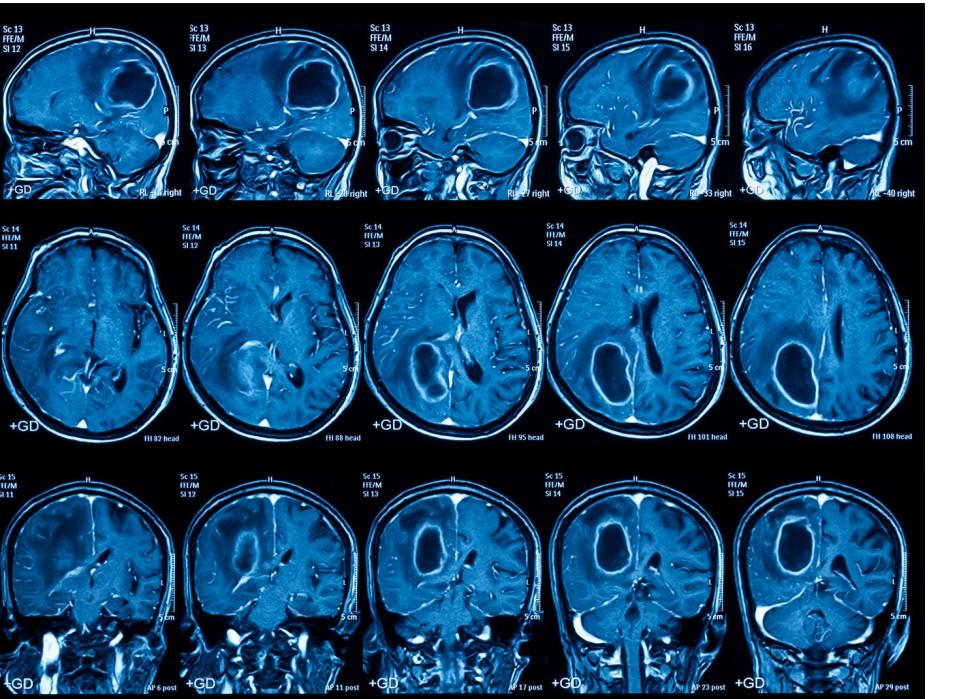
image →



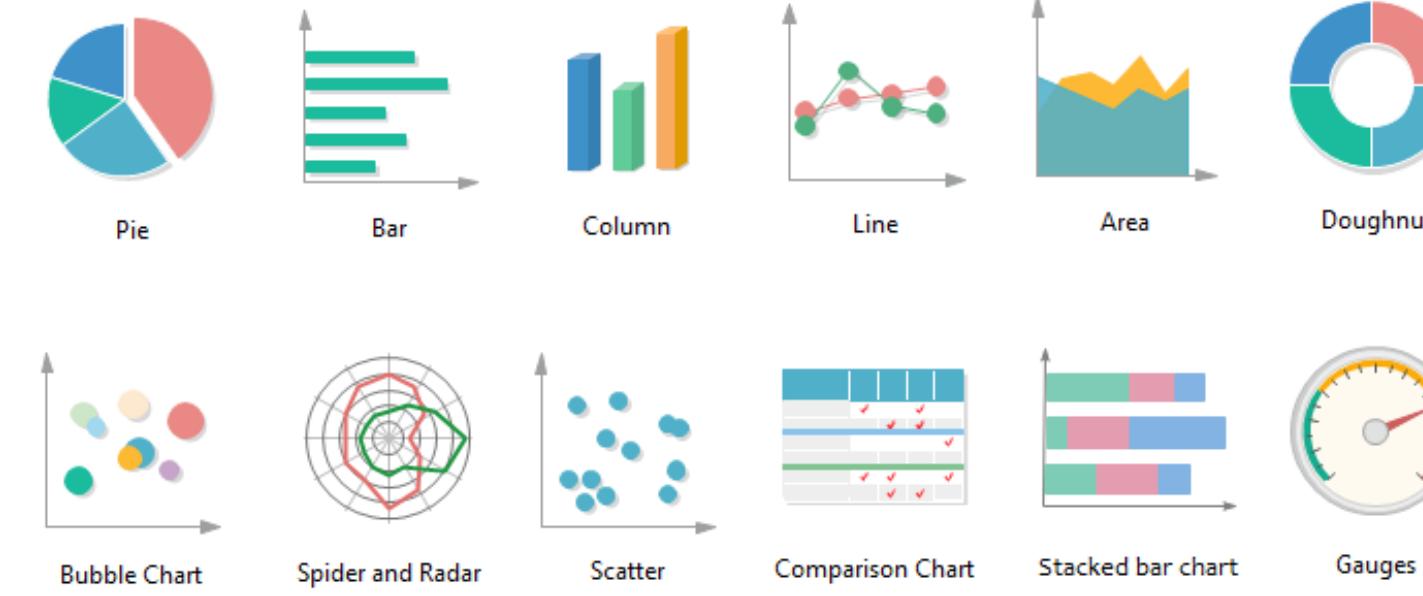
text →



→ text



medical



robot



visualizations

---

# Visual Instruction Tuning

---

**Haotian Liu<sup>1\*</sup>, Chunyuan Li<sup>2\*</sup>, Qingsyang Wu<sup>3</sup>, Yong Jae Lee<sup>1</sup>**

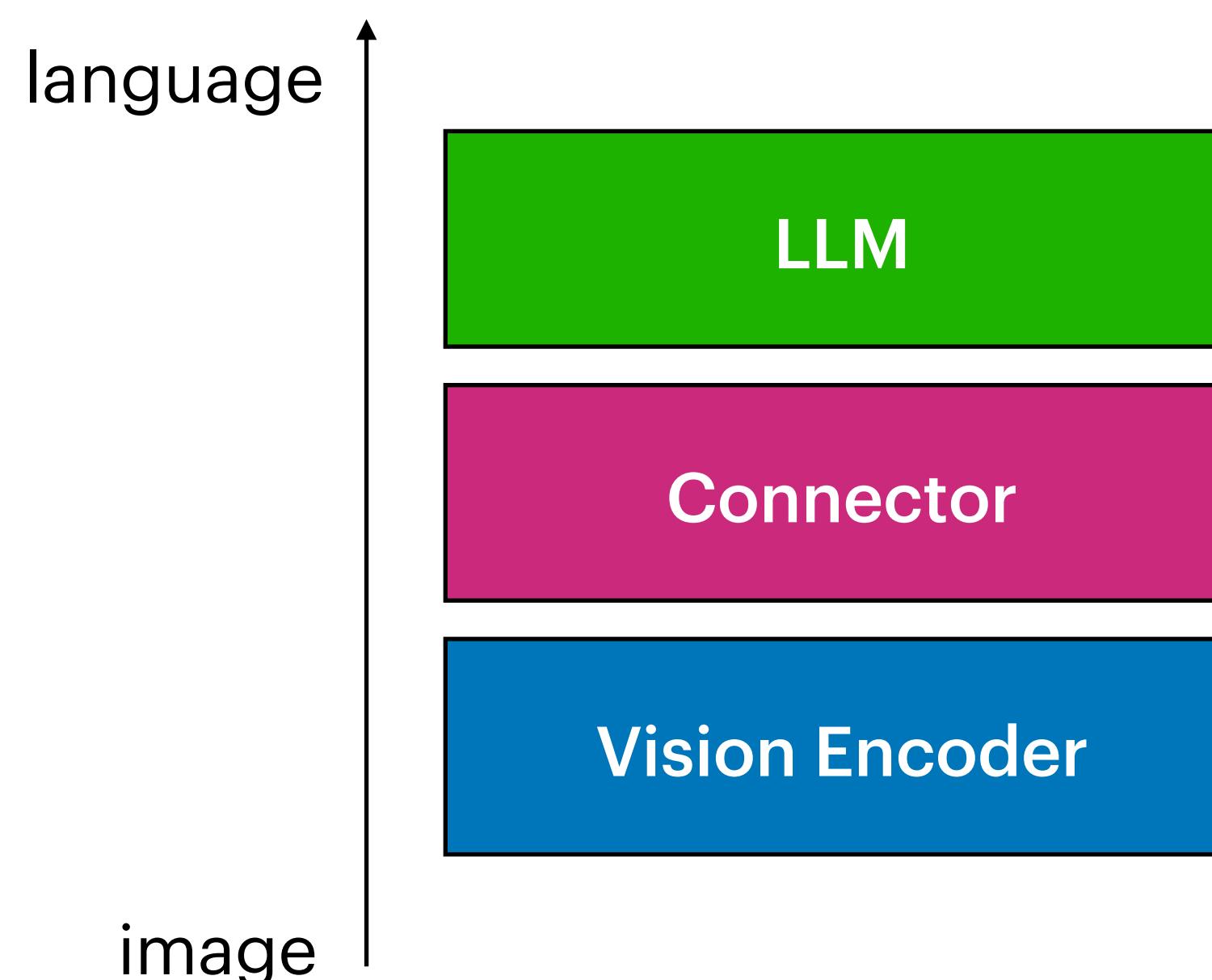
<sup>1</sup>University of Wisconsin–Madison    <sup>2</sup>Microsoft Research    <sup>3</sup>Columbia University

<https://llava-vl.github.io>

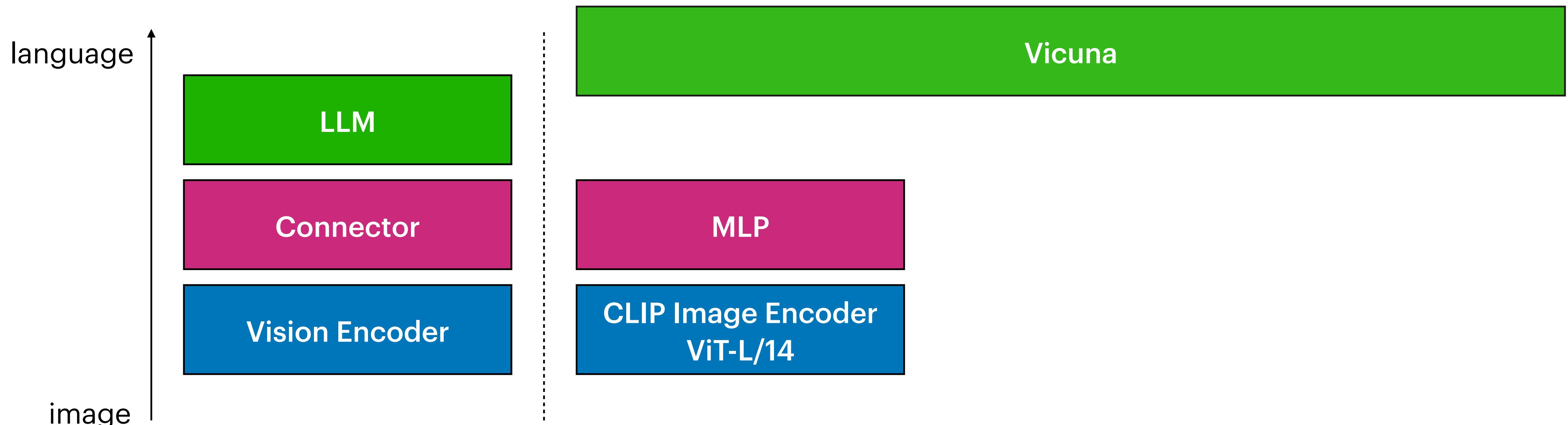
## Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has been shown to improve zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. We present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce **LLaVA: Large Language and Vision Assistant**, an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general-purpose visual and language understanding. To facilitate future research on visual instruction following, we construct two evaluation benchmarks with diverse and challenging application-oriented tasks. Our experiments show that LLaVA demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model, and code publicly available.

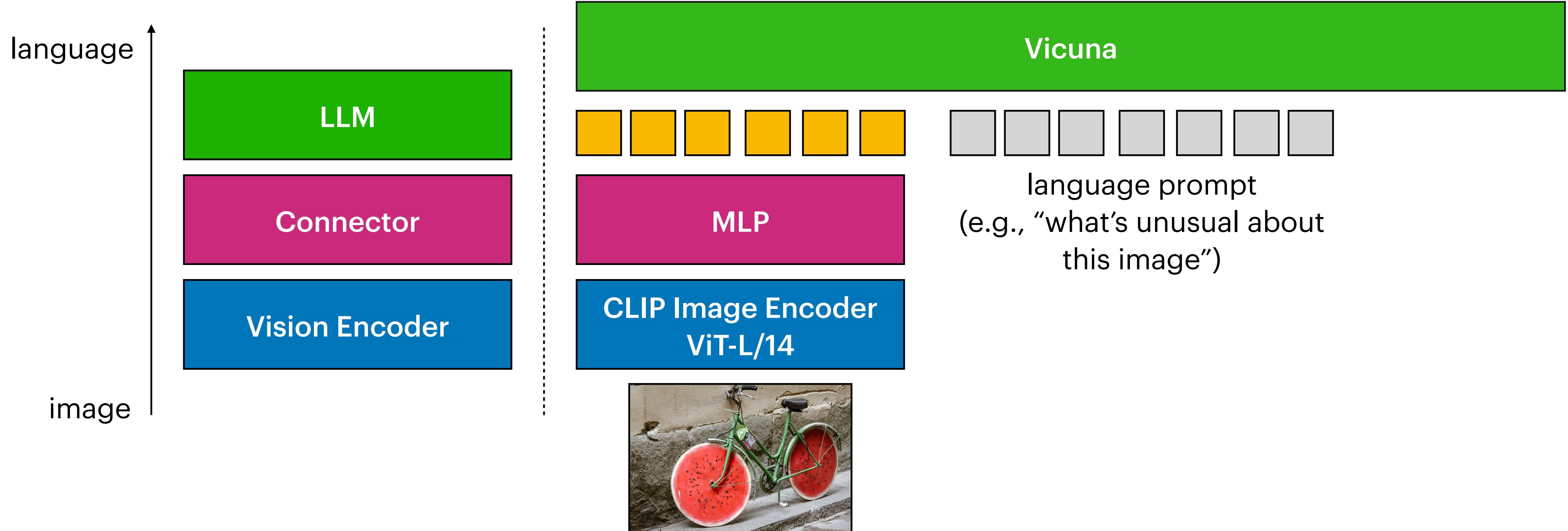
# LLaVA



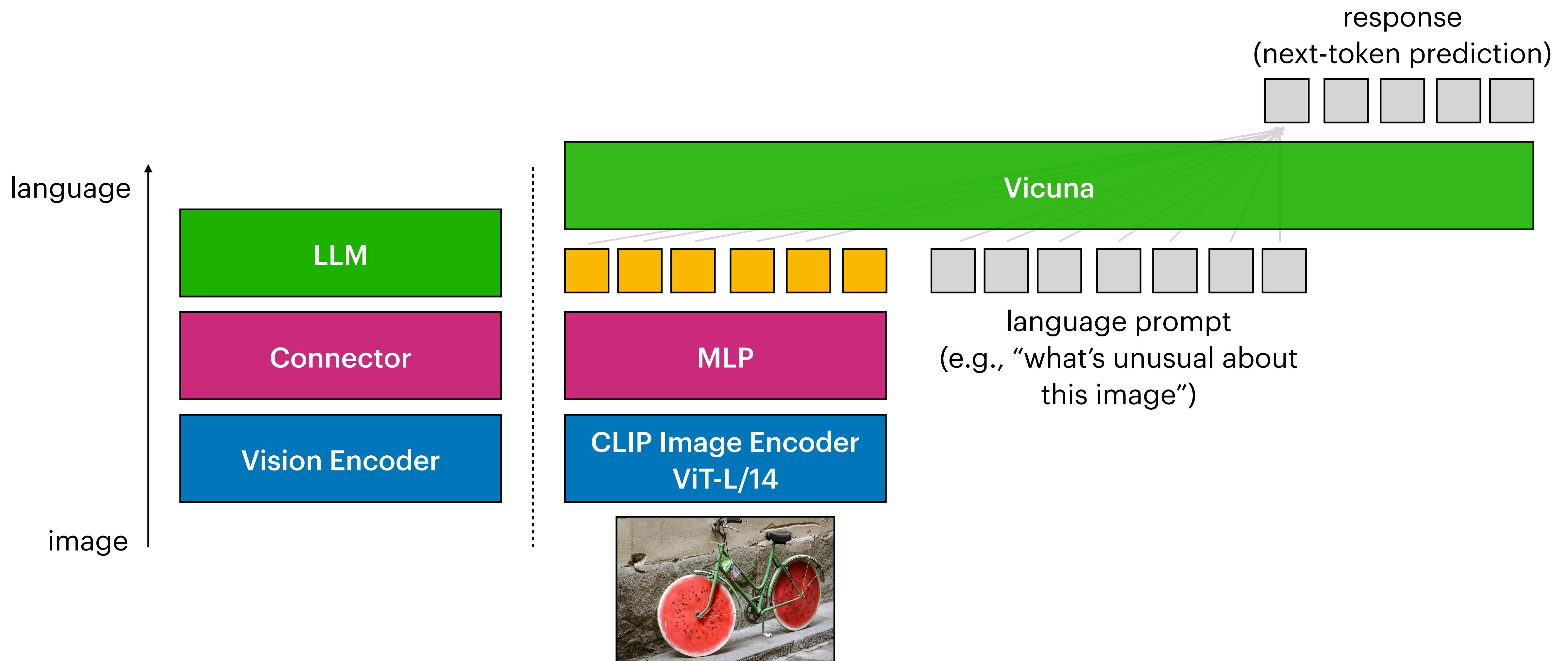
# LLaVA



# LLaVA



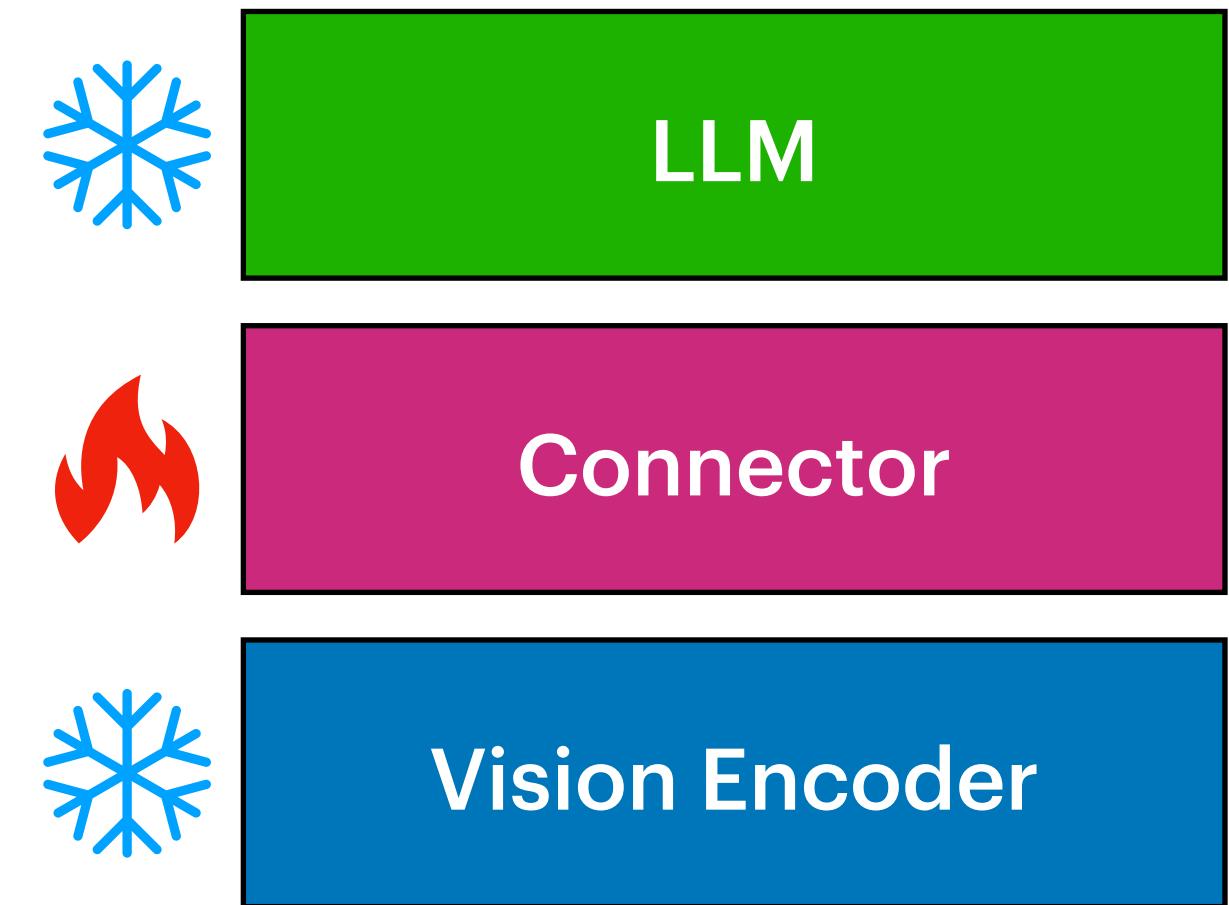
# LLaVA



# LLaVA

- Stage #1: feature alignment
  - fix V & LLM, update MLP
  - 595K image-text pairs repurposed for instruction following
- Stage #2: instruction tuning
  - fix V, update MLP & LLM
  - 158K visual instruction data

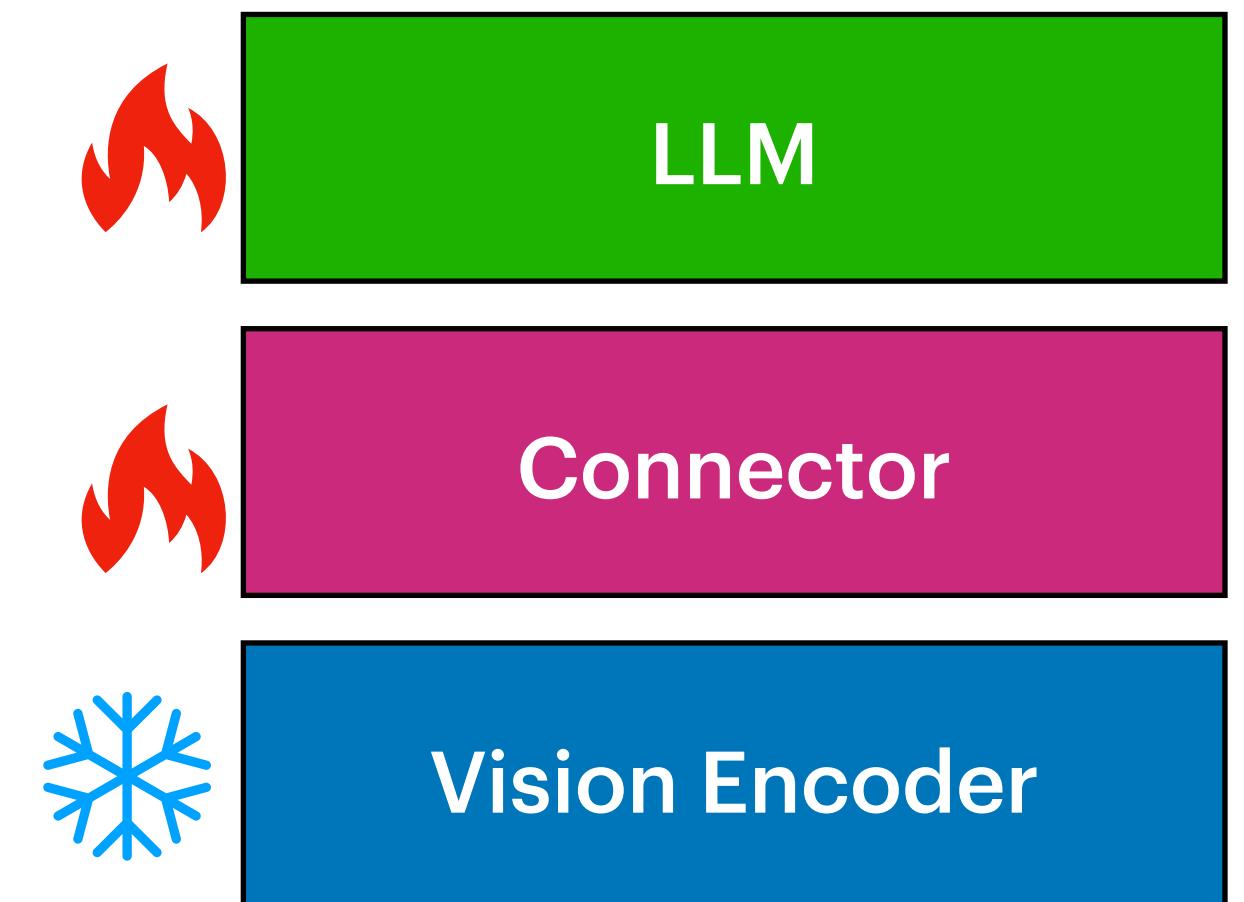
"In this way, the image features can be aligned with the pre-trained LLM word embedding. This stage can be understood as training a compatible visual tokenizer for the frozen LLM."



- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

# LLaVA

- Stage #1: feature alignment
  - fix V & LLM, update MLP
  - 595K image-text pairs repurposed for instruction following
- Stage #2: instruction tuning
  - fix V, update MLP & LLM
  - 158K visual instruction data (distilled from GPT-4)



#### Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.  
Luggage surrounds a vehicle in an underground parking area  
People try to fit all of their luggage in an SUV.



Input to GPT-4 (text only)

The sport utility vehicle is parked in the public garage, being packed for a trip  
Some people with luggage near a van that is transporting it.

#### Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Output from GPT-4

#### Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

#### Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

#### Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

# LLaVA

- Capabilities
  - visual recognition & multi-turn
  - identify atypical things
  - relate to LLM knowledge
  - generalization

# LLaVA

- Capabilities
  - **visual recognition & multi-turn**
  - identify atypical things
  - relate to LLM knowledge
  - generalization



User:

**What are the meals** that I can cook with these?

LLaVA:

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the **strawberries, blueberries, and carrots**. ...

User:

I would like to do the fruit salad. **Show me the recipe.**

LLaVA:

Certainly! Here's a simple fruit salad recipe using the items in your refrigerator.

Ingredients:

- **2 cups of fresh strawberries, sliced**
- **1 cup fresh blueberries**

...

# LLaVA

- Capabilities
  - visual recognition & multi-turn
  - **identify atypical things**
  - relate to LLM knowledge
  - generalization



User:

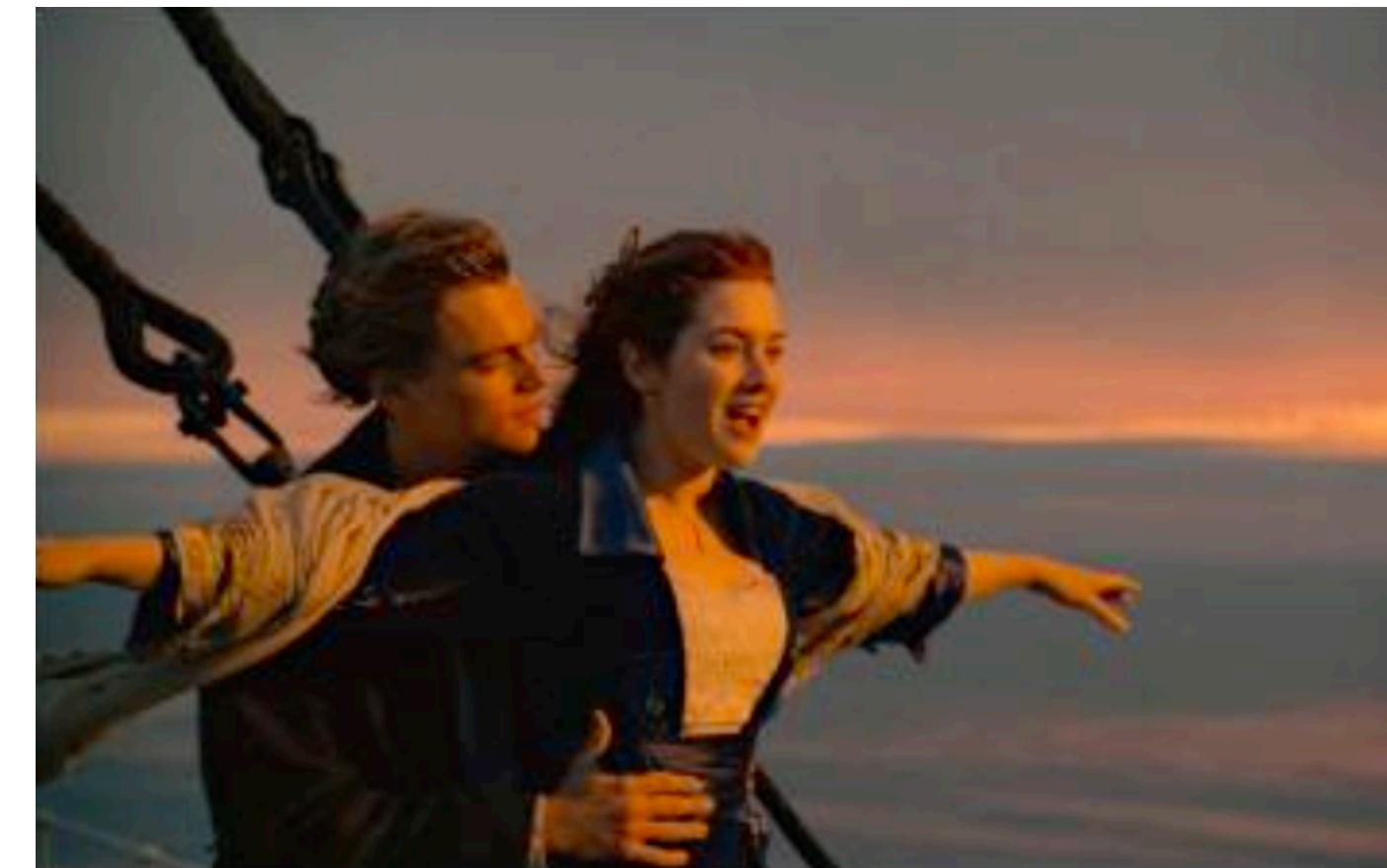
What is **unusual about this image?**

LLaVA:

The unusual aspect of this image is **a man ironing clothes on the back of a minivan or van**. This is not a typical place to perform this activity, as **one would usually iron clothes in a more stationary and safe location**, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

# LLaVA

- Capabilities
  - visual recognition & multi-turn
  - identify atypical things
  - **relate to LLM knowledge**
  - generalization



User:

What's the **ending of this movie?**

LLaVA:

In the image, a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. They appear to be enjoying a romantic moment together. The scene suggests that **they could be depicted as the characters Jack and Rose from the popular movie Titanic**. In the movie, their romance blossoms aboard the fictional ocean liner Titanic, with the pair ultimately sharing a passionate kiss as the ship sinks into the ocean. The ending of the movie is known for its emotional and heartfelt scenes, where the characters are forced to say goodbye to each other as they are separated by the sinking ship. The image evokes a sense of romance and nostalgia, paying tribute to the iconic love story portrayed in the movie Titanic.

# LLaVA

- Capabilities
  - visual recognition & multi-turn
  - identify atypical things
  - relate to LLM knowledge
- **generalization**



User:

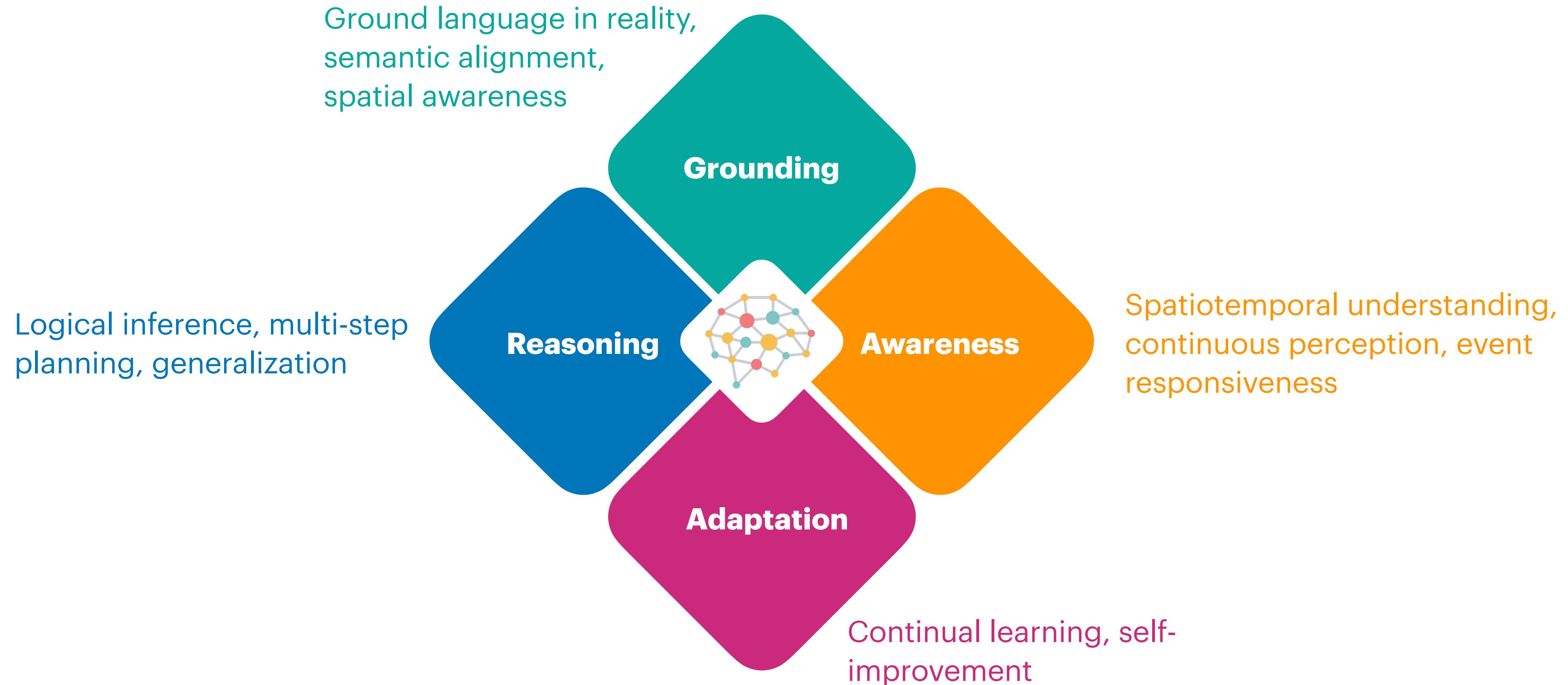
What is the **name of the man** in the photo?

LLaVA:

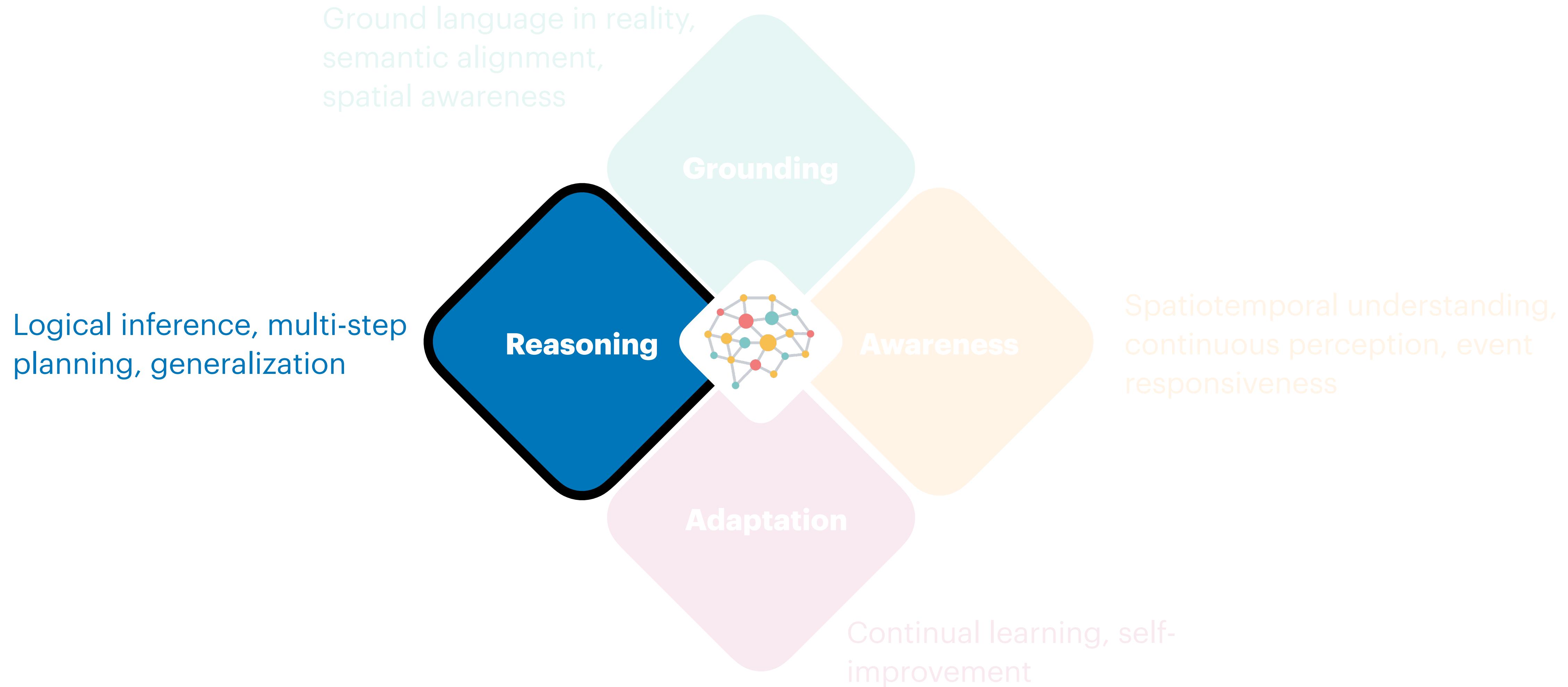
The name of the man in the photo is **Elon Musk**.

(Elon Musk never appears during visual instruction tuning. This ability may come from the pre-trained vision encoder, meaning that the LLM generalizes to unseen visual concepts.)

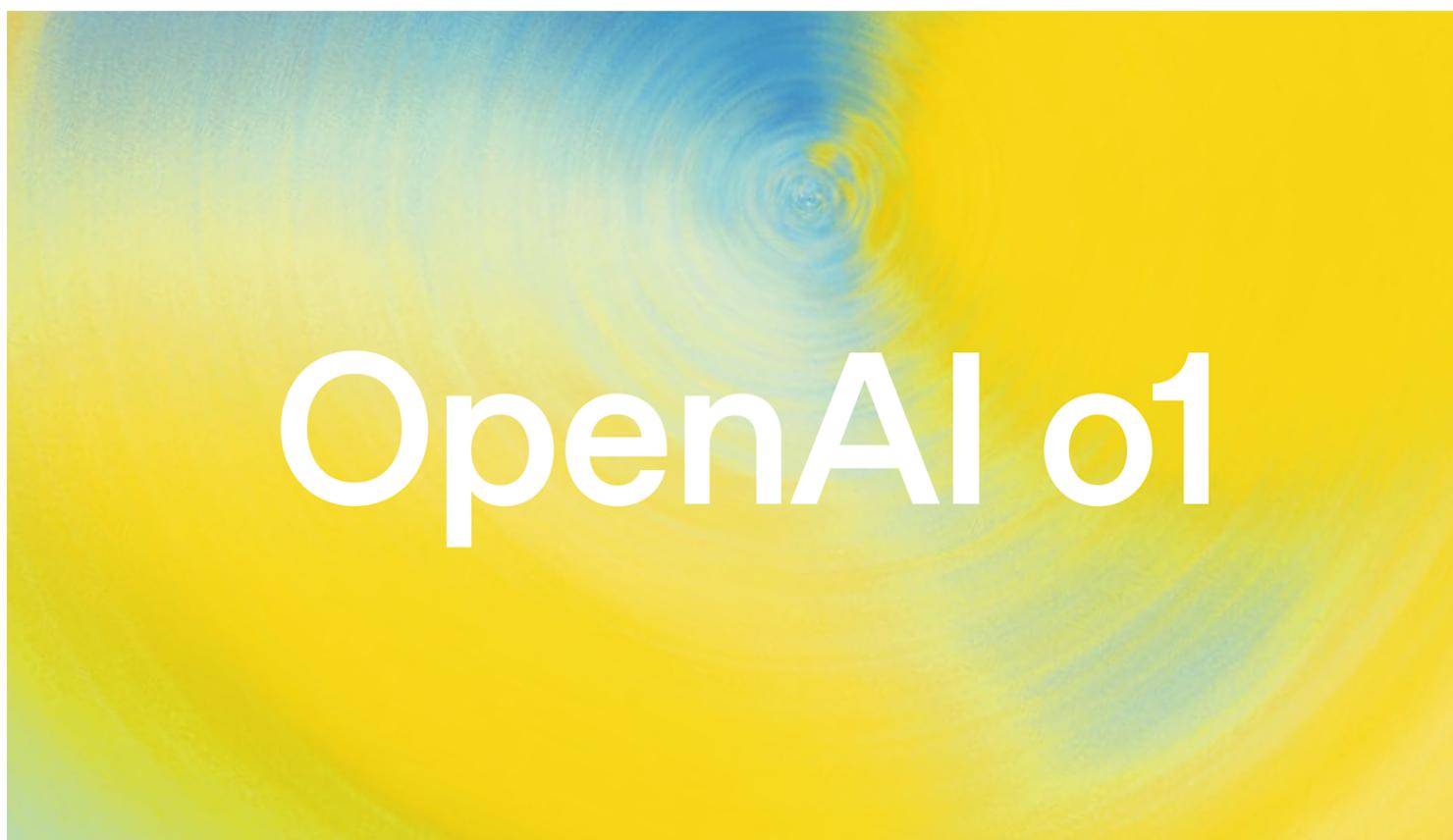
# Cornerstones of multimodal agents



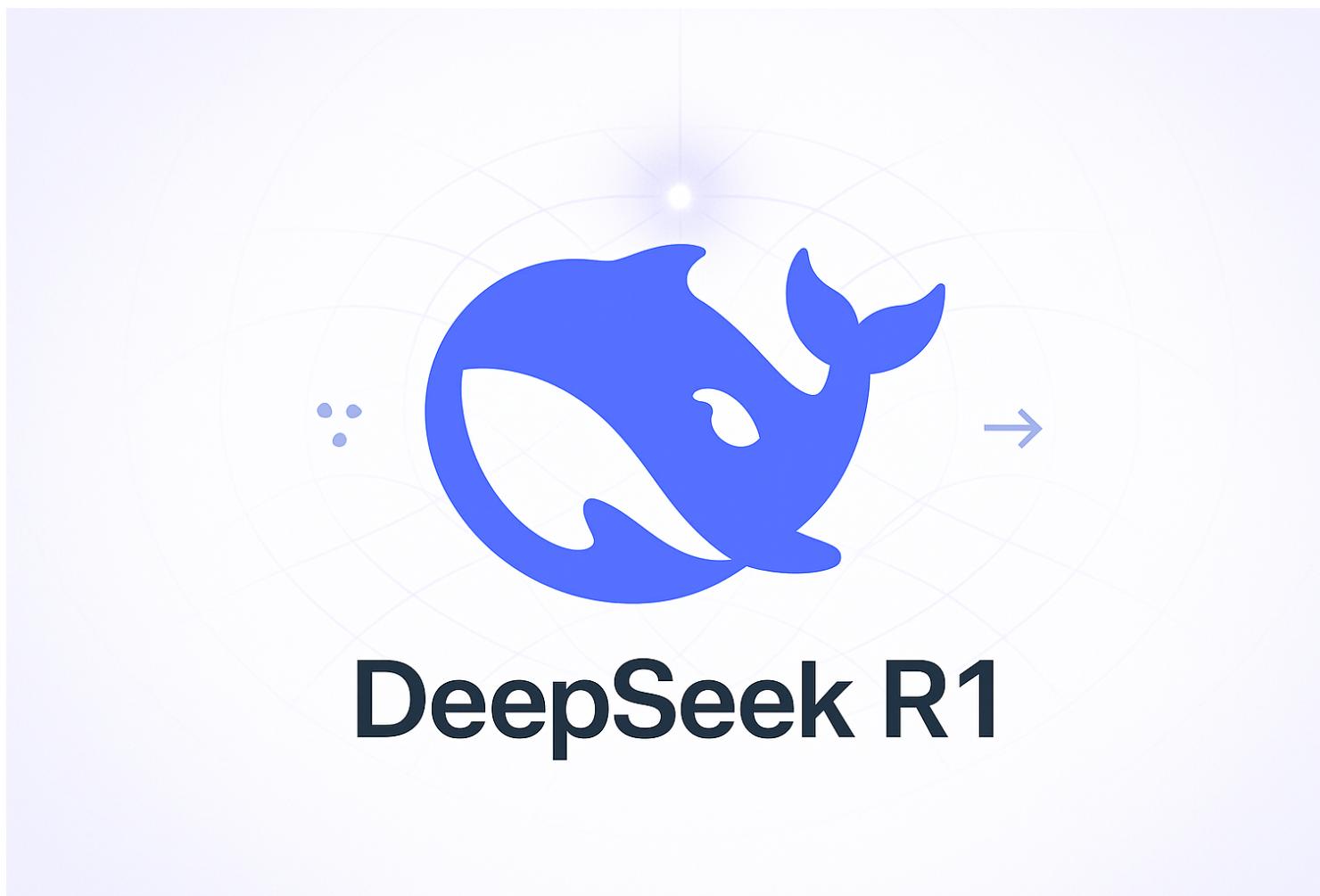
# Cornerstones of multimodal agents



# Deep thinking



release in Sep 2024



release in Jan 2025

"...to spend more time **thinking** before they respond"  
"...excels at **STEM**, especially math and coding"

8 apples 5 oranges 25 bananas 8 grapes 15  
strawberries 23 watermelons 1 apple 18  
raspberries 5 lemons 25 kiwis 15 peaches 21  
blueberries

Thought for 5 seconds ^

**Deciphering the message**  
I'm connecting the numbers to letters, examining fruits, and looking for any hidden patterns. Progress is being made as I explore these possibilities.

**Mapping positions**  
I'm thinking through the numbers as positions in the alphabet to link them with the names of fruits, like using their 8th letter or alternate methods for mapping.

HEY HOW ARE YOU

🔊 🔍 🔍 🔍 🔍 🔍

Twitter @SmokeAwayyy

# A typical multimodal reasoning paradigm

## 1. Pre-training

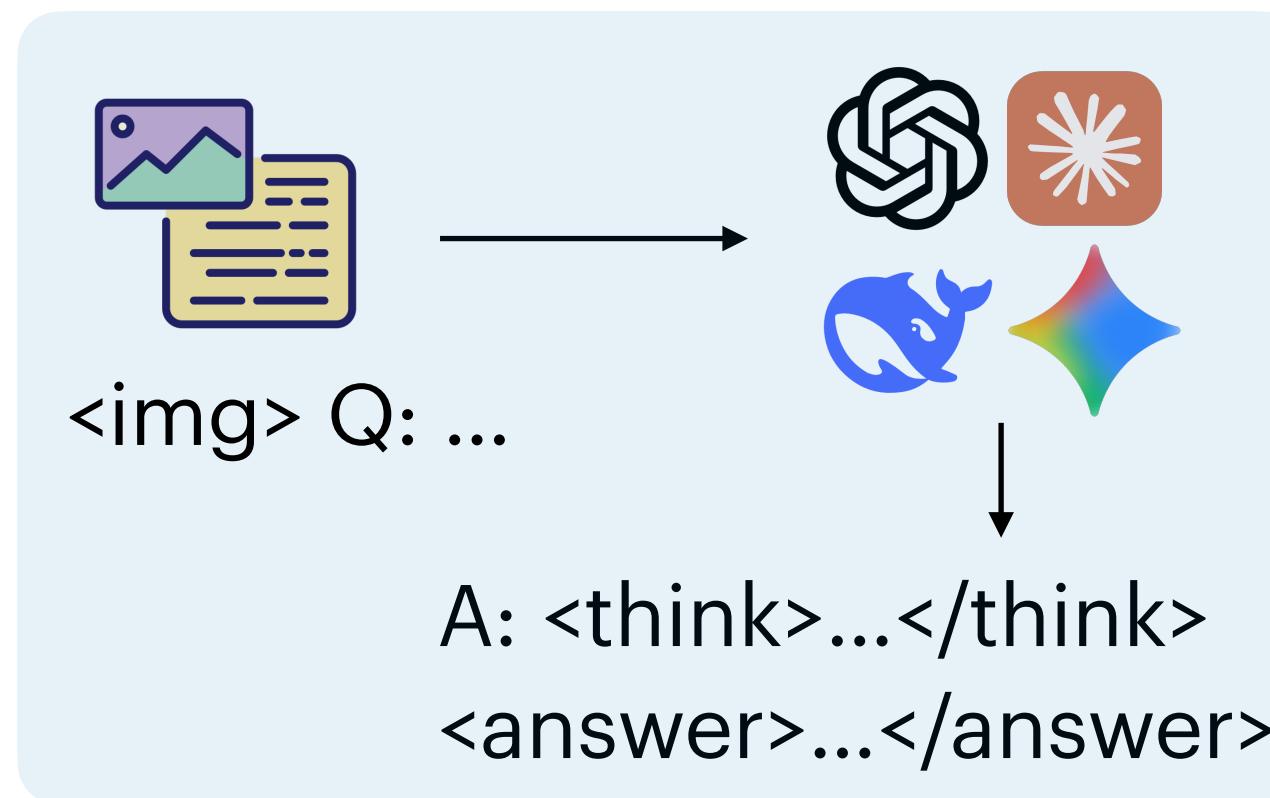
Pre-train VLM on image instruction data



SFT

## 2. SFT

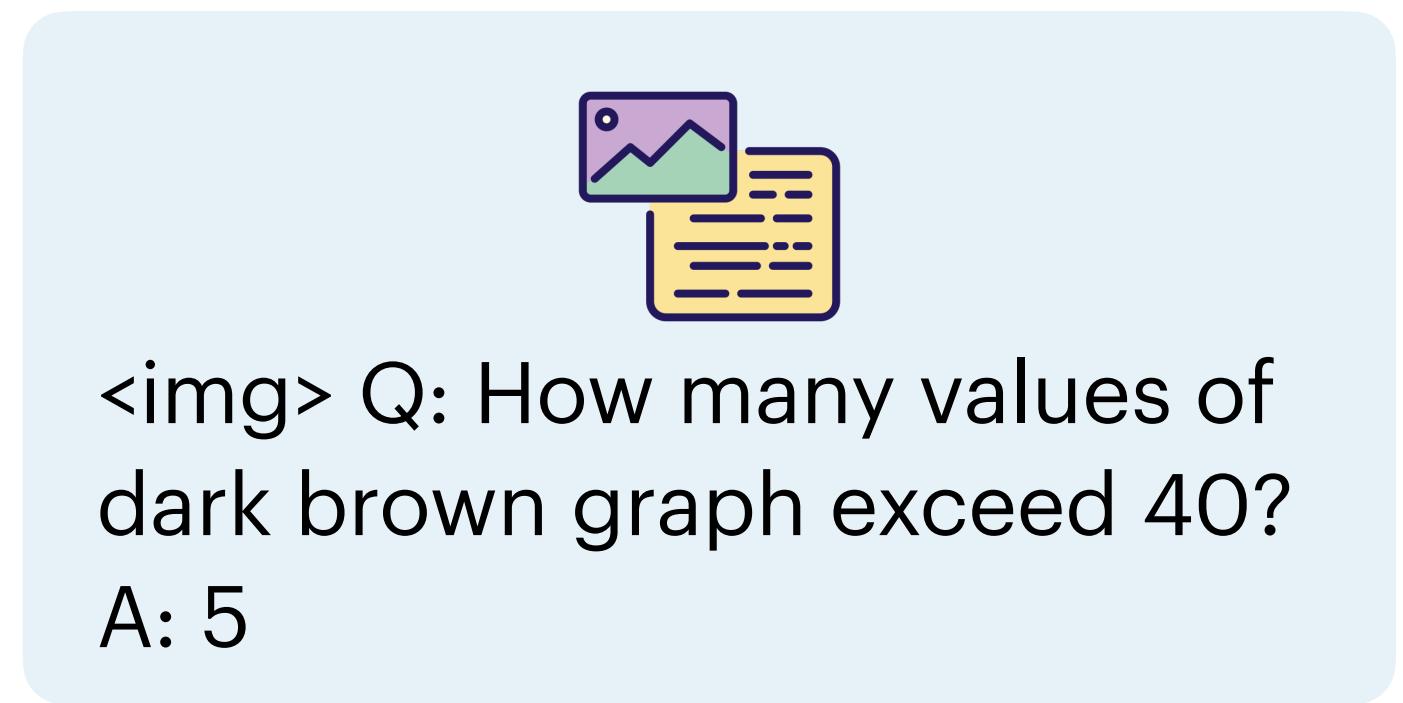
Fine-tune VLM on distilled chain-of-thought data



SFT

## 3. RL

Fine-tune VLM via rule-based RL on VQA data



GRPO

<think>...</think>  
<answer>5</answer>



<think>...</think>  
<answer>3</answer>



<think>blahblahblah  
lahblahblahblahblah  
blahblahblahblahbla  
hblahblahblahblahbl  
ahblahblah</think>...

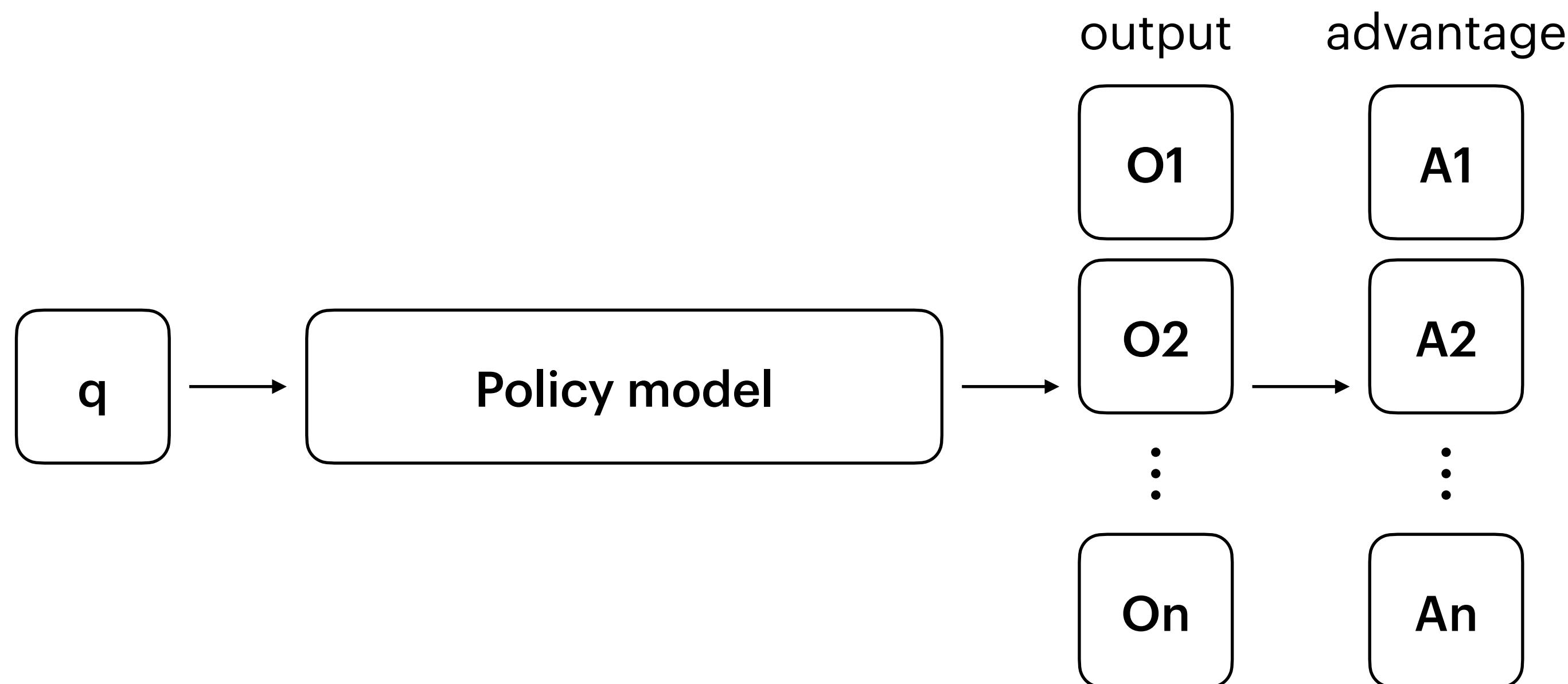


<think>blahblahblah  
lahblahblah</think>...



# GRPO

## Group Relative Policy Optimization

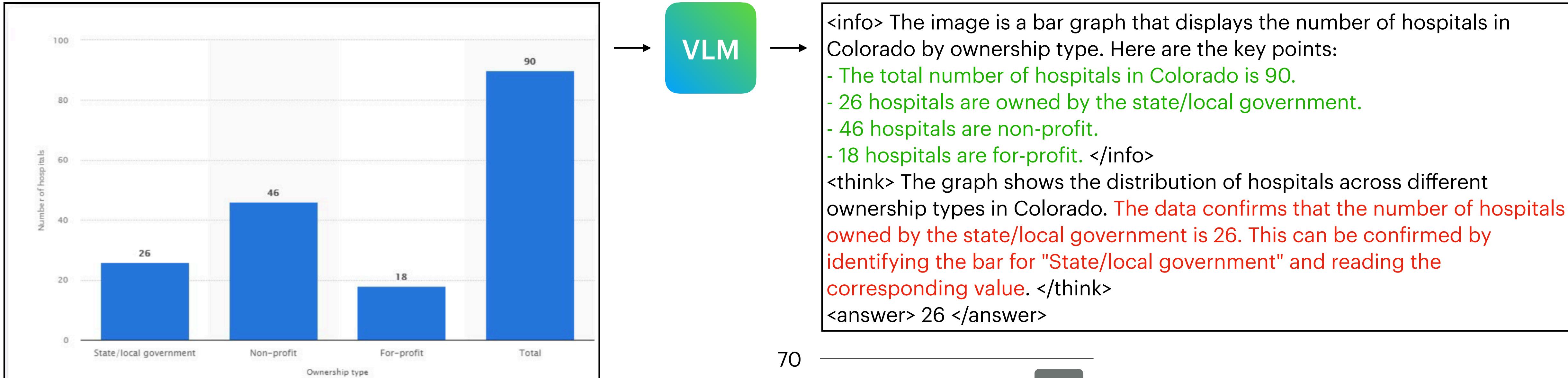


$$A_i = \frac{r_i - \text{mean}(r_1, \dots, r_n)}{\text{std}(r_1, \dots, r_n)}$$

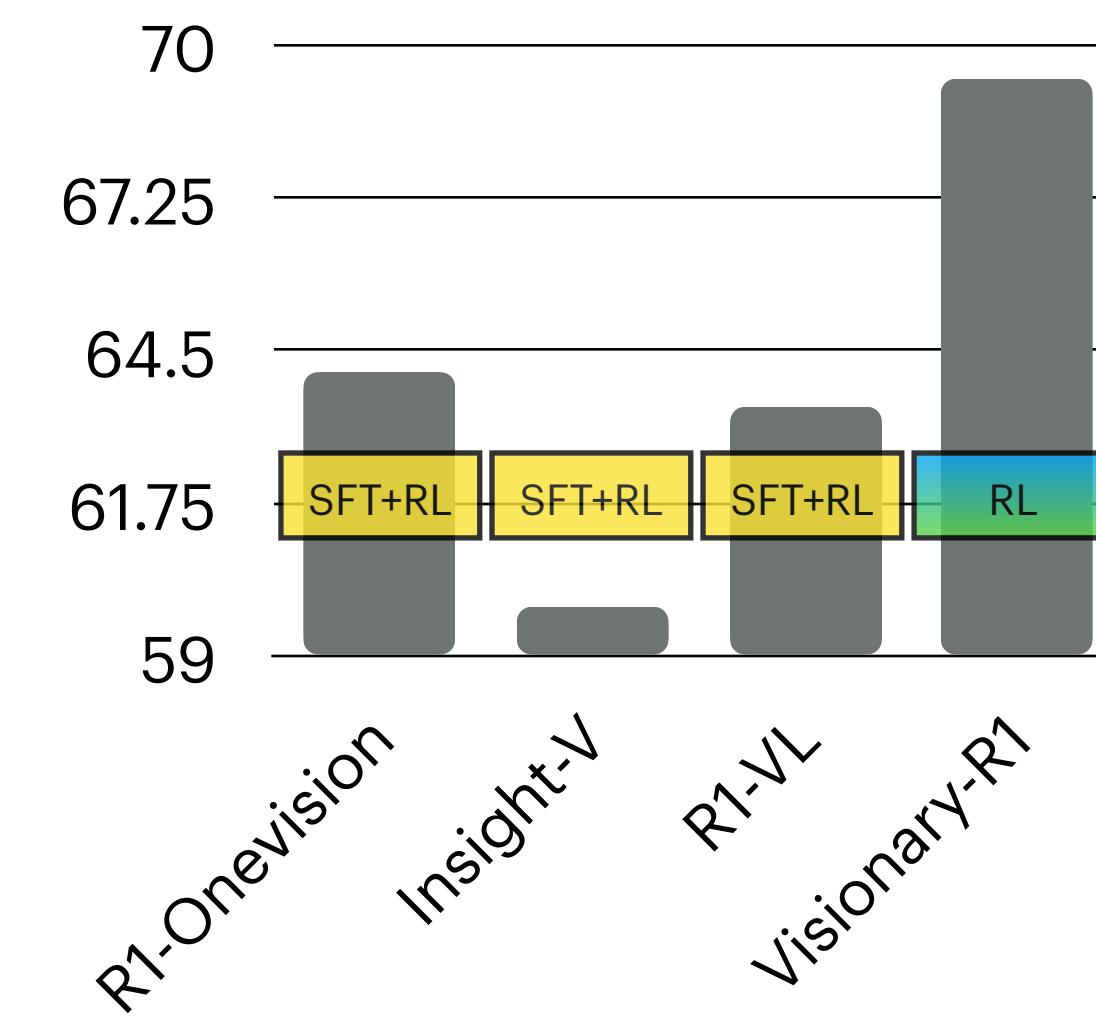
$$r_i = r_{acc}(o_i) + r_{format}(o_i)$$

<think>...</think>  
<answer>...</answer>

# Key idea: interpret image first, then reason

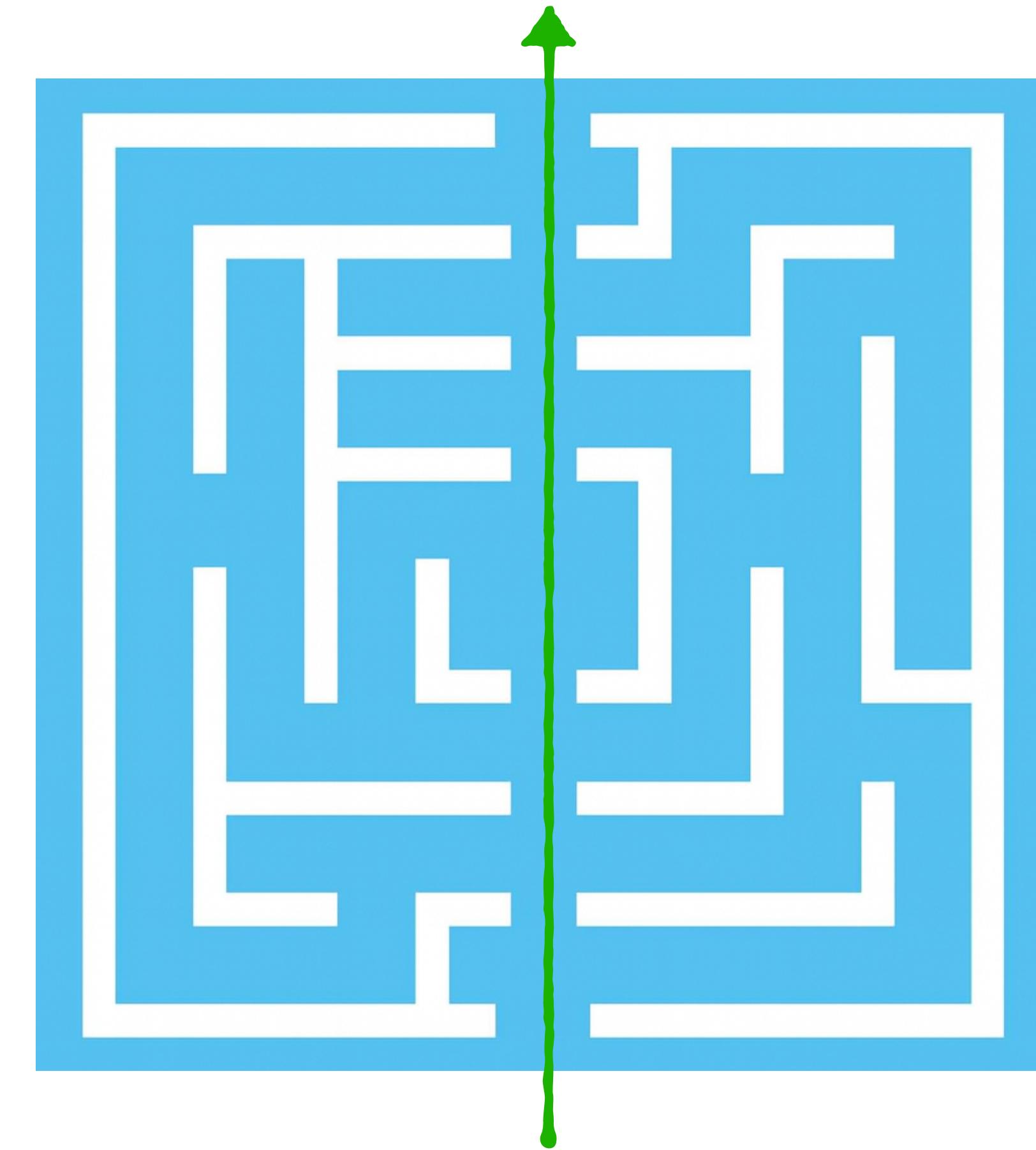
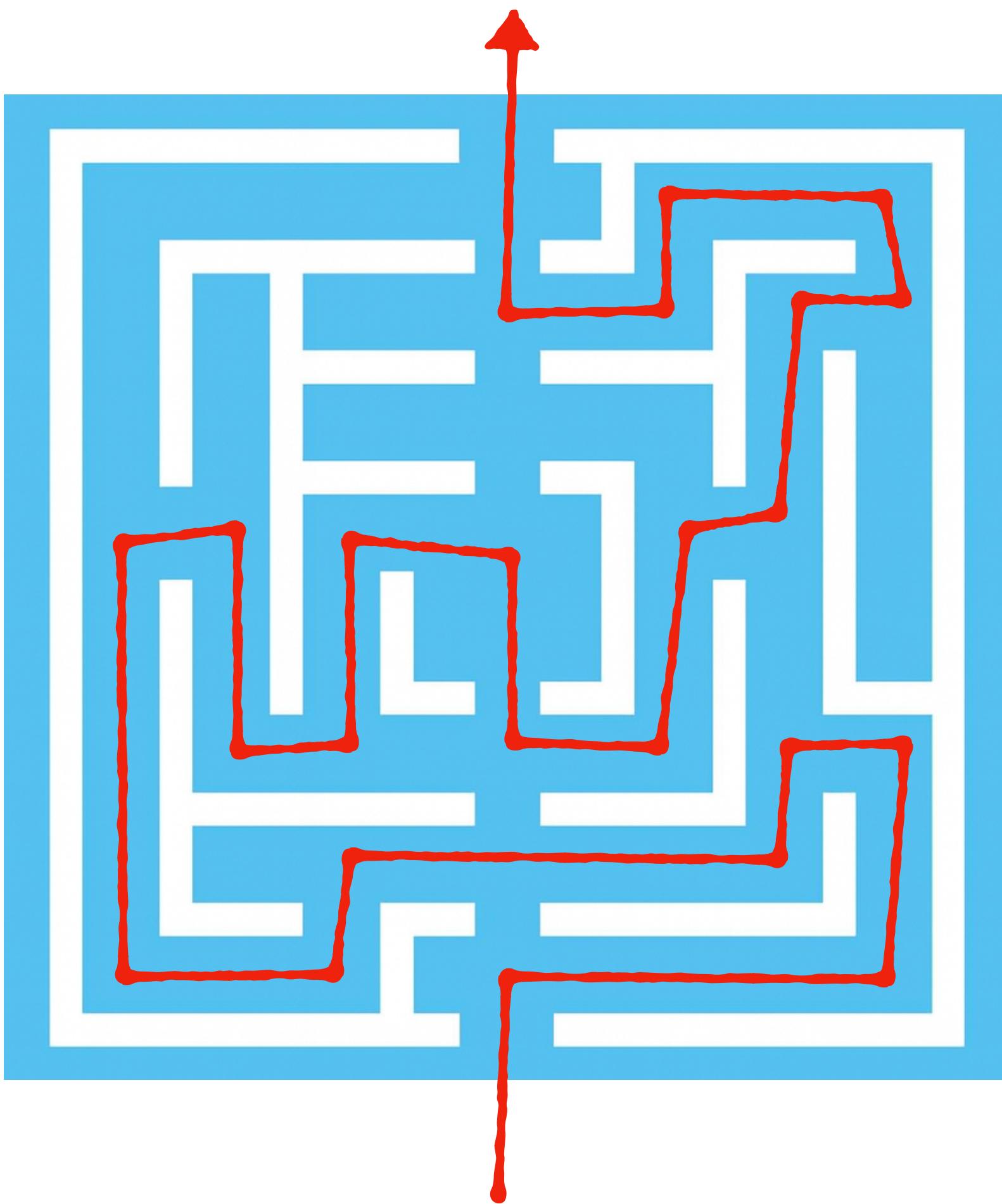


Q: How many hospitals in Colorado were owned by the state or local government in 2019?



Visionary-R1 achieves SOTA on MathVista, despite using RL only

# Do we really need to go *long* every time?



'A lifetime's worth of wisdom'  
Steven D. Levitt, co-author of *Freakonomics*

## The International Bestseller

# Thinking, Fast and Slow



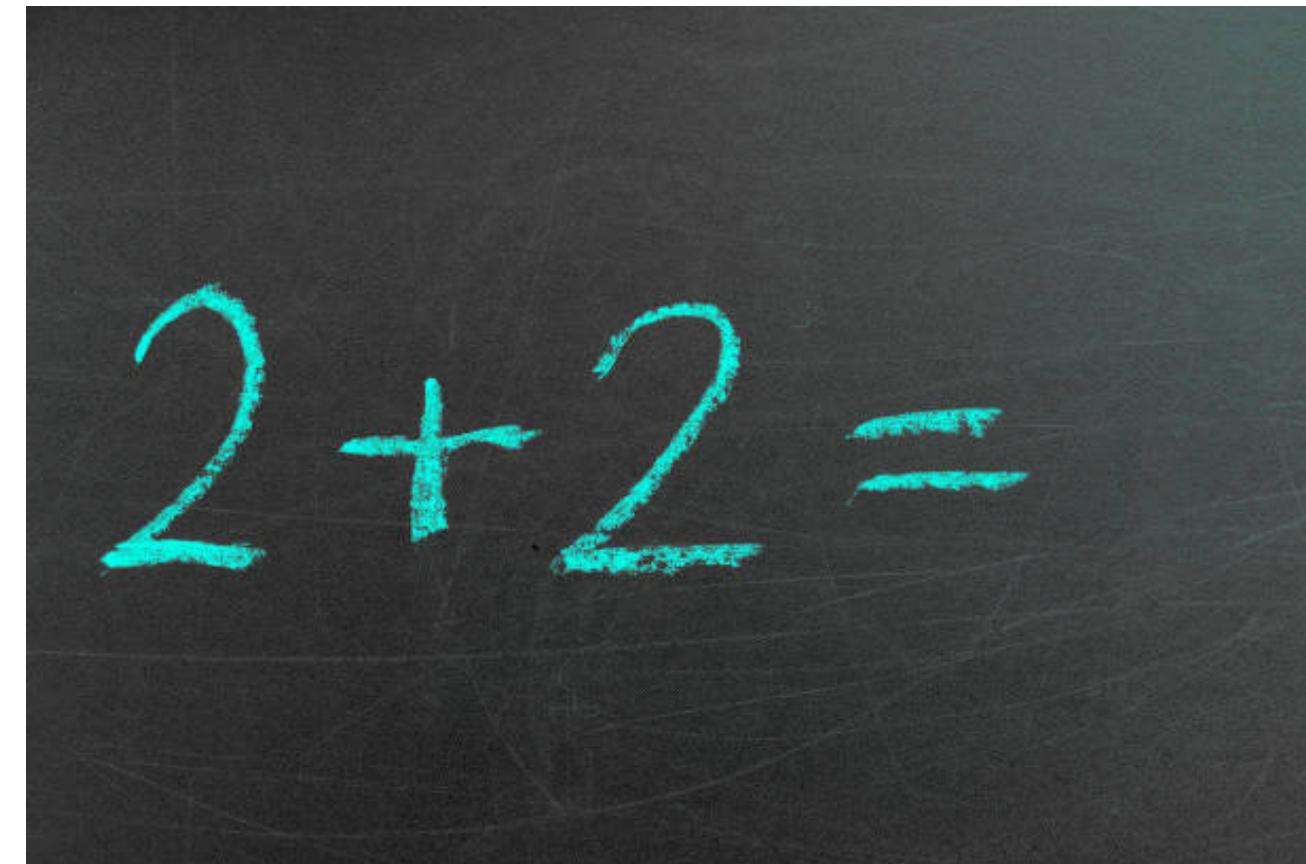
Daniel Kahneman

Winner of the Nobel Prize



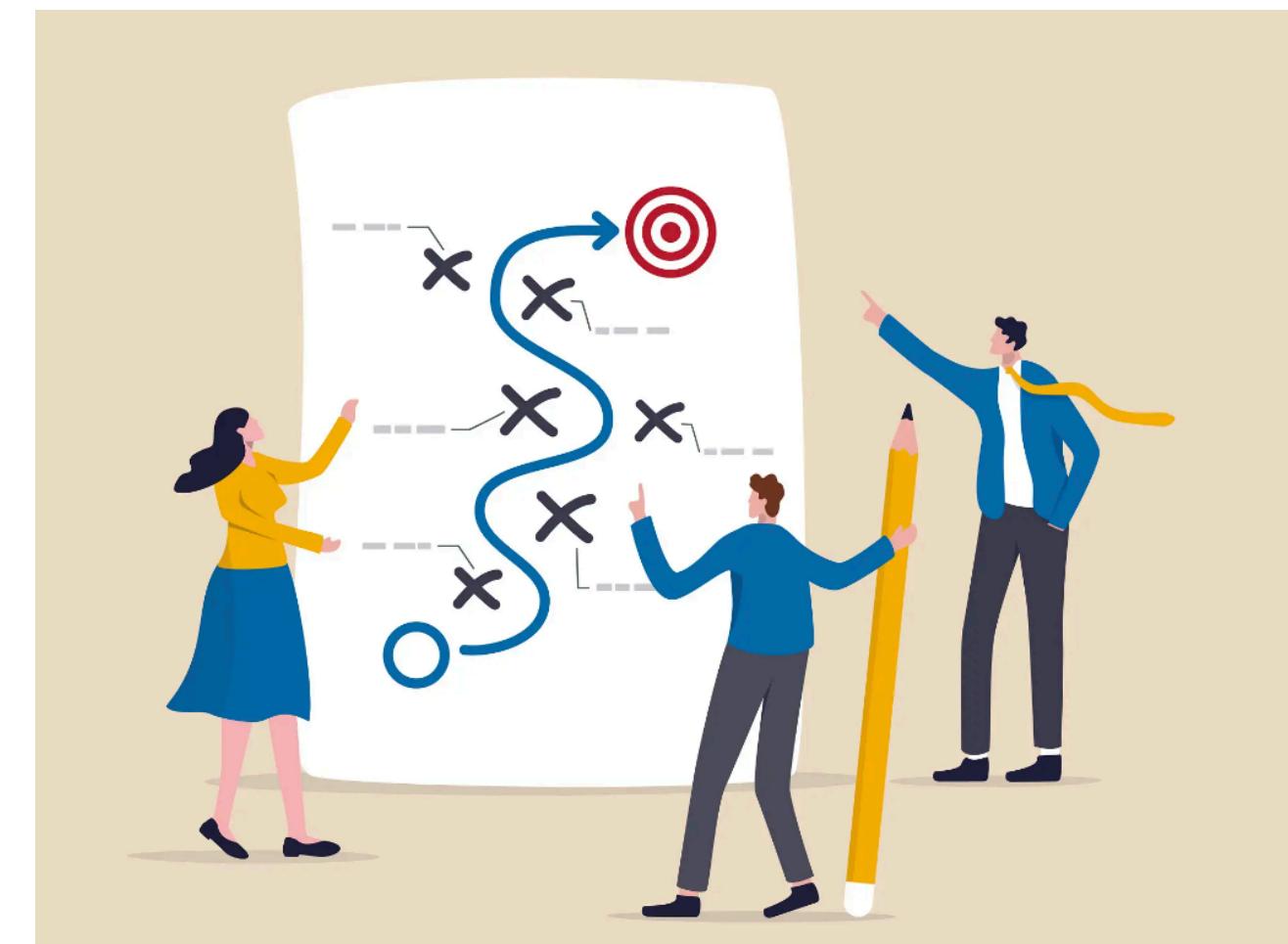
- **System 1 thinking**

- fast
- automatic
- intuitive



- **System 2 thinking**

- slow
- deliberate
- analytical



# Survey: which model do you prefer?



What emotion does this emoji express?  
Options: A: happy, B: sad, C: excited,  
D: angry



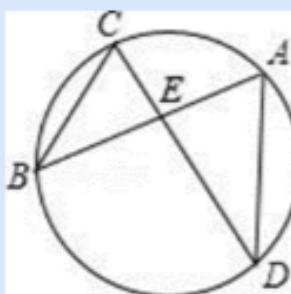
## Model 1



To determine the emotion expressed by the emoji, let's analyze its features:

1. The emoji has closed eyes, which often indicate a state of rest, sleep, or perhaps  
... Therefore, the correct answer is \boxed{B}

126 Tokens



As shown in the figure, the two chords AB and CD in the circle intersect at E,  $\angle D = 35^\circ$ ,  $\angle AEC = 105^\circ$ , then  $\angle C = 0$   
Options: A:60°, B: 70°, C:80°, D:85°



## Model 1



To solve for  $\angle C$ , we need to use the properties of angles in a circle, specifically the fact that the sum of angles in a triangle is  $180^\circ$  ...

- Step 1: Identify the given information and the relationships between the angles.  
 $\angle D = 35^\circ$ ,  $\angle AEC = 105^\circ$  ...  
Step 2: ...

Therefore, the correct answer is \boxed{B}

556 Tokens



## Model 2



Short Thinking: The emoji has closed eyes and a downturned mouth, which are typical expressions of sadness. \boxed{B}

26 Tokens



## Model 2



Long Thinking: To solve this problem, we need to use the properties of angles ... \*\*Identify the given information: ... \*\*Understand the relationship between the angles:\*\* ... Thus, the correct answer is \boxed{B}.

280 Tokens

# Survey: which model do you prefer?



What emotion does this emoji express?  
Options: A: happy, B: sad, C: excited,  
D: angry



**GRPO**

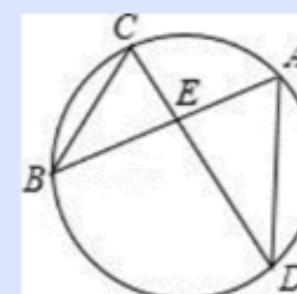


To determine the emotion expressed by the emoji, let's analyze its features:

1. The emoji has closed eyes, which often indicate a state of rest, sleep, or perhaps

... Therefore, the correct answer is \boxed{B}

126 Tokens



As shown in the figure, the two chords AB and CD in the circle intersect at E,  $\angle D = 35^\circ$ ,  $\angle AEC = 105^\circ$ , then  $\angle C = 0$   
Options: A:60°, B: 70°, C:80°, D:85°



**GRPO**



To solve for  $\angle C$ , we need to use the properties of angles in a circle, specifically the fact that the sum of angles in a triangle is  $180^\circ$  ...

Step 1: Identify the given information and the relationships between the angles.

$\angle D = 35^\circ$ ,  $\angle AEC = 105^\circ$  ...

Step 2: ...

Therefore, the correct answer is \boxed{B}

556 Tokens



**DualMindVLM**



Short Thinking: The emoji has closed eyes and a downturned mouth, which are typical expressions of sadness. \boxed{B}

26 Tokens



**DualMindVLM**



Long Thinking: To solve this problem, we need to use the properties of angles ... \*\*Identify the given information: ... \*\*Understand the relationship between the angles:\*\* ... Thus, the correct answer is \boxed{B}.

280 Tokens



## DualMindVLM

1. auto-label QA as easy or hard
2. RL: think fast for easy, think slow for hard



easy



hard

$|\text{response}| < 100$

$|\text{response}| > 200$

# Model output length as an indicator



What emotion does this emoji express?  
Options: A: happy, B: sad, C: excited,  
D: angry



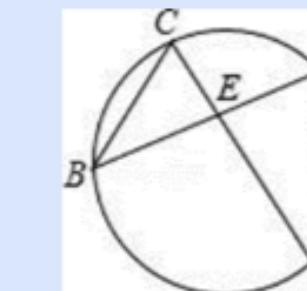
## Base Model



The emoji in the image shows a face with closed eyes and a slight frown, which typically represents a sad or disappointed emotion.

\boxed{B: sad}

34 Tokens



As shown in the figure, the two chords AB and CD in the circle intersect at E,  $\angle D = 35^\circ$ ,  $\angle AEC = 105^\circ$ , then  $\angle C = 0$   
Options: A:60°, B: 70°, C:80°, D:85°



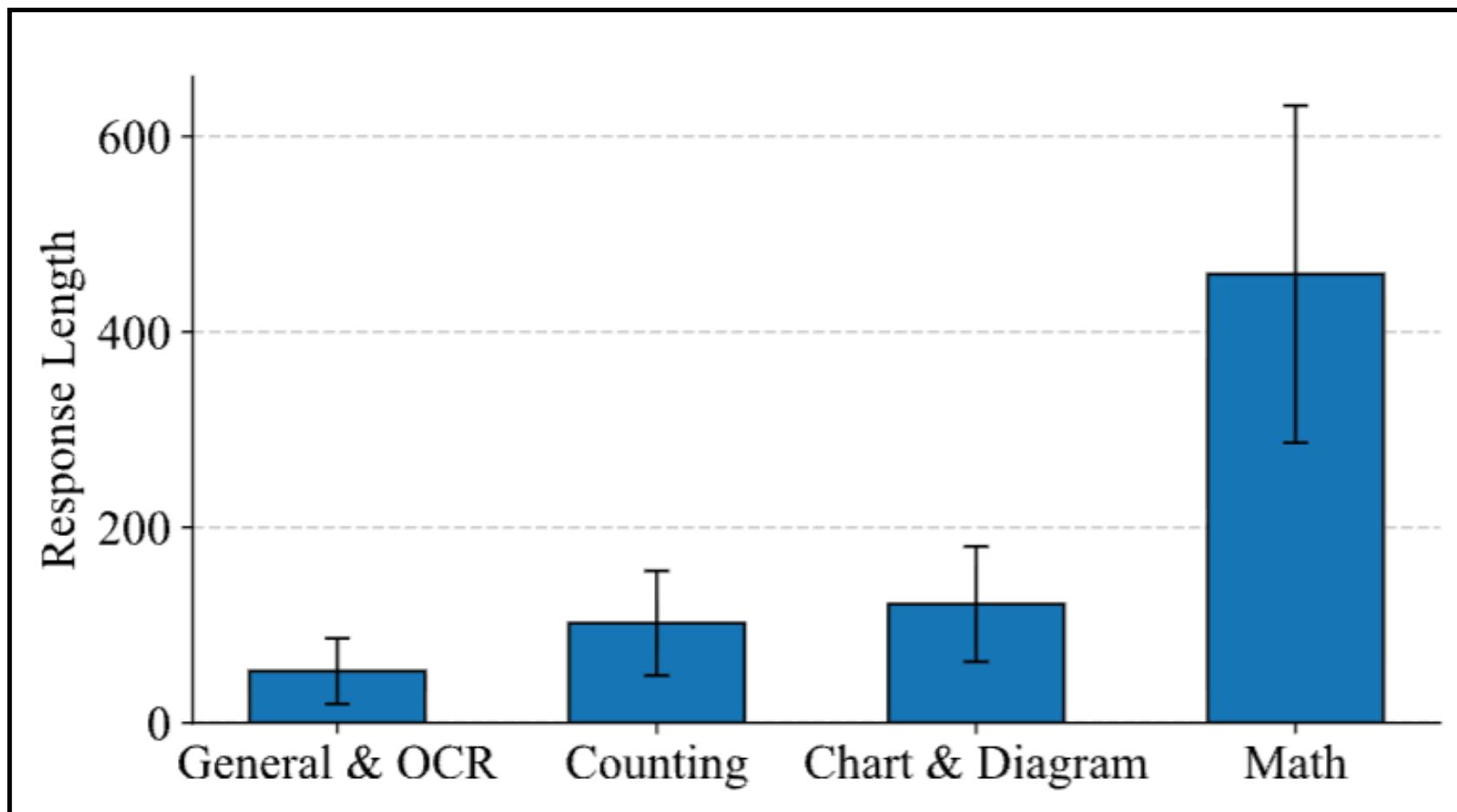
## Base Model



To solve for the measure of  $\angle C$  in the given circle, we need to use the properties of angles in a circle, specifically the fact that the measure of an angle formed by ... Here are the steps: ...

So, the measure of  $\angle C$  is \boxed{75°}

361 Tokens

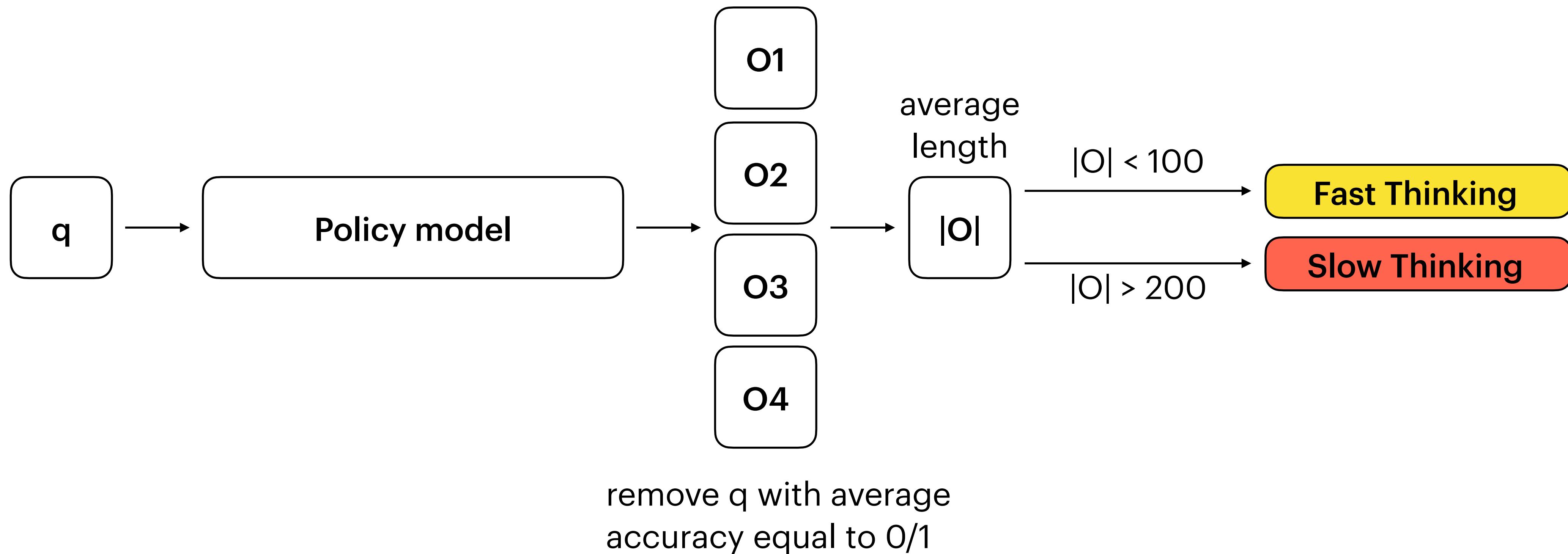


The longer the length, the harder the question.

The shorter the length, the easier the question.

# DualMindVLM

## Stage 1: Thinking mode auto-labeling



# DualMindVLM

## Stage 2: Learning dual-mode thinking

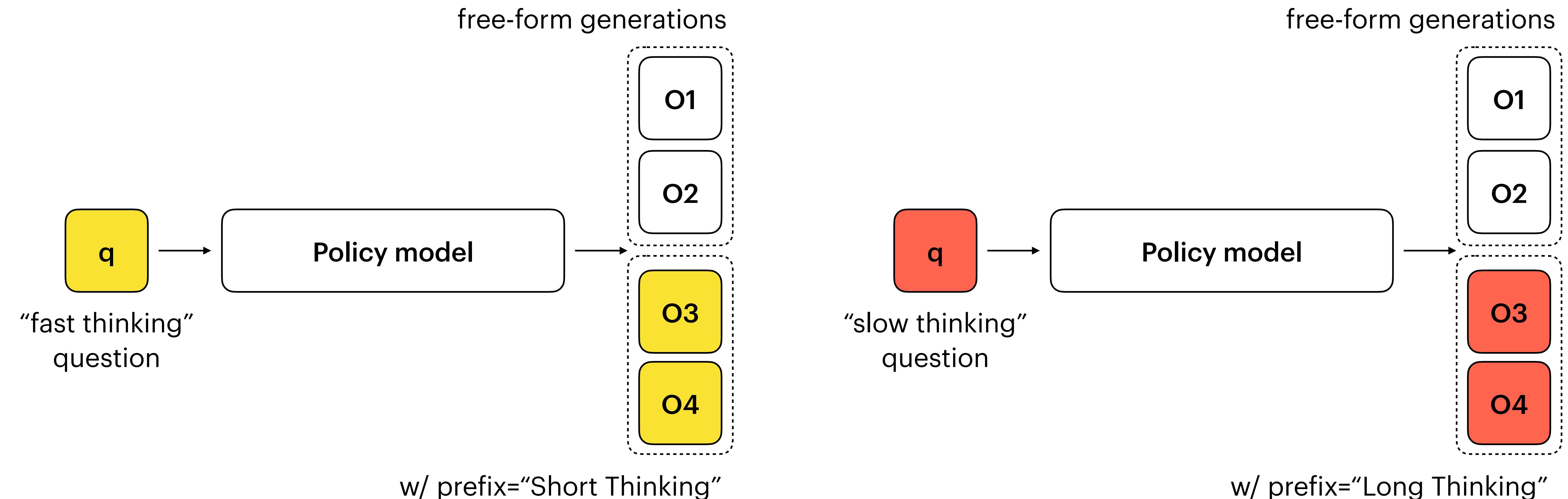
### System Prompt

You are a Vision-Language Model answering questions about images. Follow these rules strictly:

1. Judge the length of reasoning needed.
  - Short: start with "Short Thinking:".
  - Long: start with "Long Thinking:".
2. "Short Thinking:" give a concise thinking process which is sufficient to answer the question, then provide the final answer.
3. "Long Thinking:" give a structured reasoning process of the question and the image, including question analysis, visual details description, self-verification and then provide the final answer.
4. The final answer **MUST BE** put in \boxed{ }.

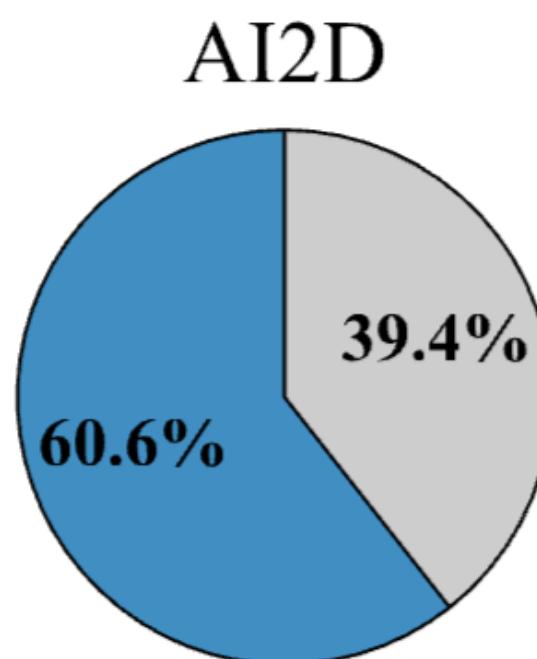
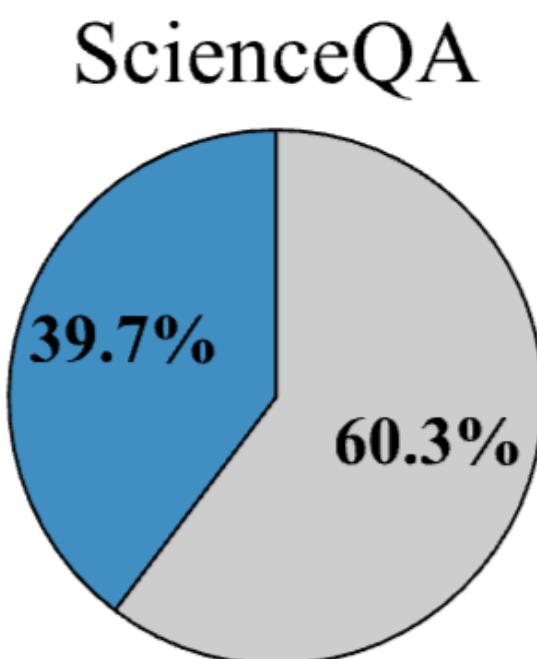
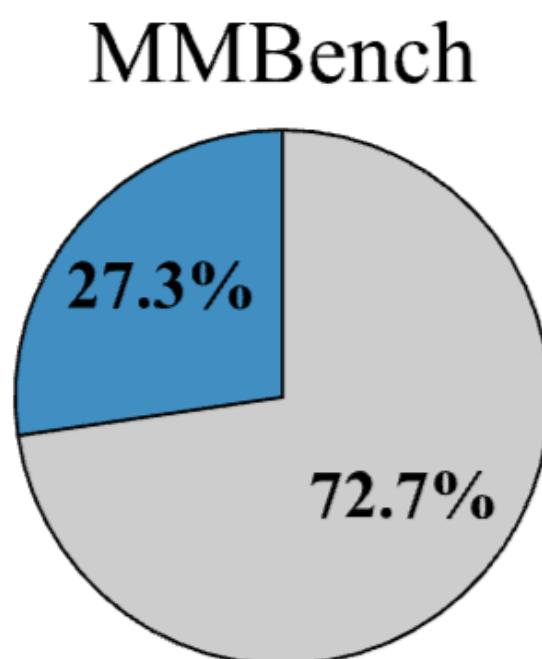
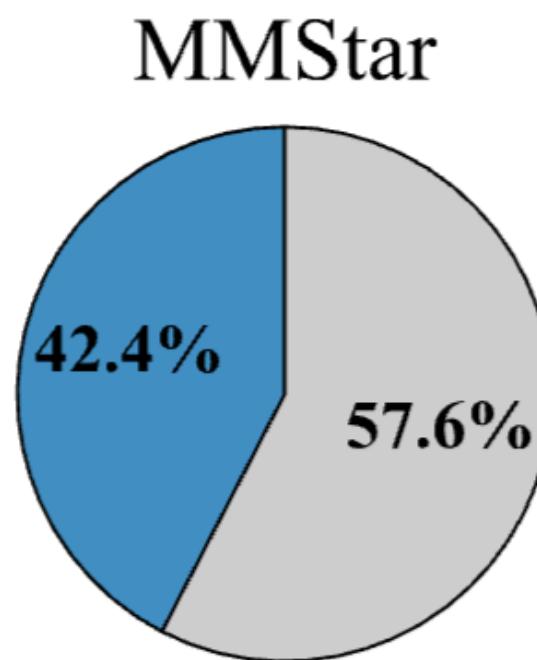
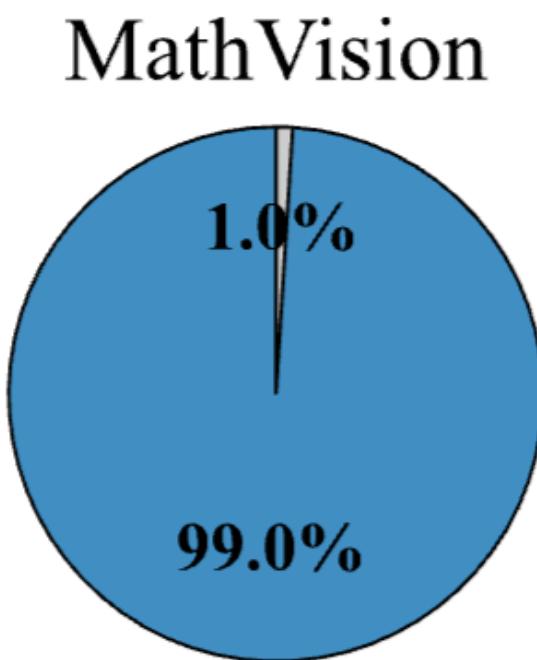
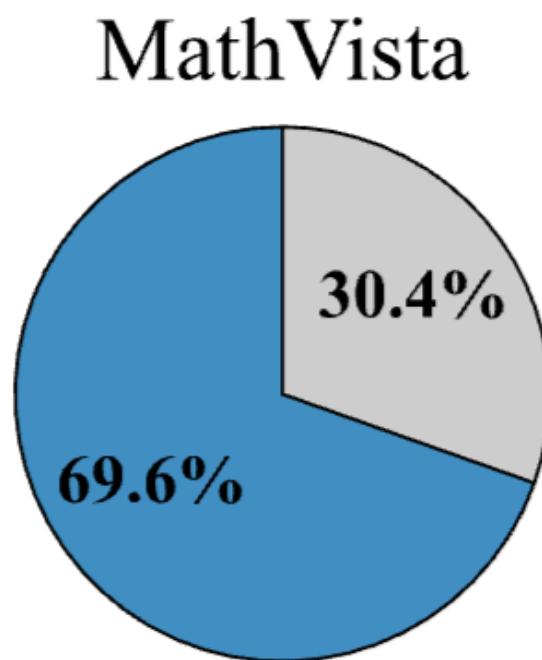
# DualMindVLM

## Stage 2: Learning dual-mode thinking



Model	Size	Strategy	MathVista		MathVision		MMStar		MMBench		ScienceQA		AI2D		Avg	
			acc ↑	len ↓	acc ↑	len ↓	acc ↑	len ↓	acc ↑	len ↓	acc ↑	len ↓	acc ↑	len ↓	acc ↑	len ↓
Qwen2.5-VL	7B	-	68.2	205	25.1	511	63.9	155	83.0	73	84.0	156	80.8	145	67.5	208
LLaVA-CoT	11B	SFT	54.8	350	-	-	57.6	464	-	-	-	-	-	-	-	-
R1-Onevision	7B	SFT+RL	64.1	279	29.9	560	-	-	-	-	-	-	-	-	-	-
R1-VL	7B	SFT+RL	63.5	263	24.7	363	60.0	221	-	-	-	-	-	-	-	-
OpenVLThinker	7B	SFT+RL	72.3	242	25.9	<b>326</b>	63.3	200	87.5	177	82.2	171	83.2	160	69.1	213
MM-Eureka	7B	RL	73.0	252	26.9	612	64.1	246	87.3	159	83.5	202	83.5	207	69.7	280
ThinkLite	7B	RL	75.1	247	28.5	599	65.0	175	<b>88.7</b>	113	-	-	83.6	168	-	-
VL-Rethinker	7B	RL	74.9	268	<b>32.3</b>	566	64.9	231	87.6	201	85.5	205	82.4	226	71.3	283
DualMindVLM	7B	RL	<b>75.6</b>	<b>184</b>	30.2	<b>446</b>	<b>65.3</b>	<b>121</b>	88.3	<b>69</b>	<b>87.2</b>	<b>98</b>	<b>83.8</b>	<b>104</b>	<b>71.7</b>	<b>170</b>

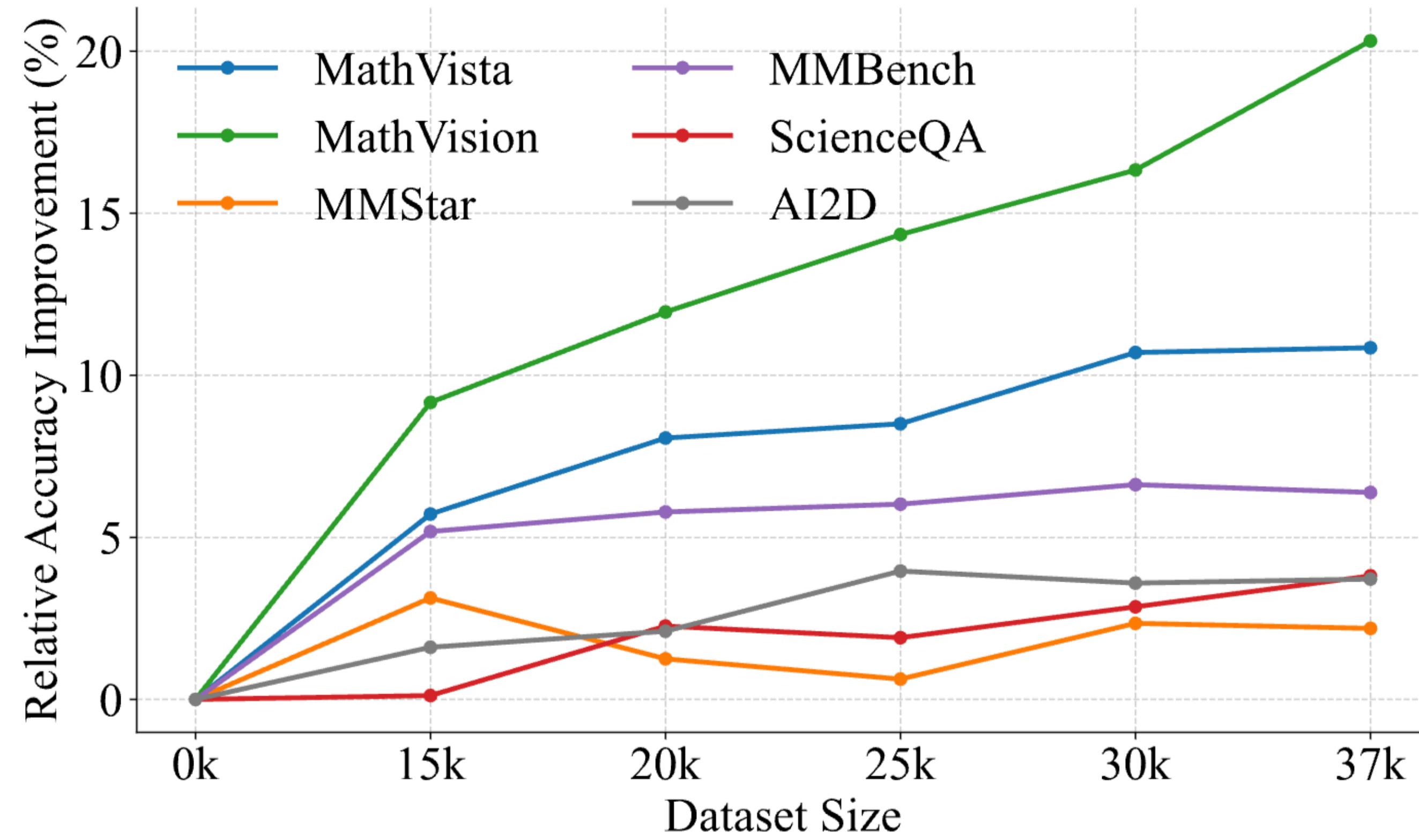
SOTA performance & high token efficiency



Calculate the ratios of slow and fast thinking modes.

Prefers **slow** thinking mode for challenging problems like math.

Relatively **balanced** mode selection on other benchmarks.



Gradually increase the training data.

Scale is important to challenging problems like **math**.

Scale has **limited** impact on perceptual/scientific tasks.

# Application — Analyze driving



Question: What color is the first car parking on the right?

DualMindVLM:

**Short Thinking:** The first car parking on the right is white. **(13 tokens)**

Question: Identify potential hazards and propose mitigations.

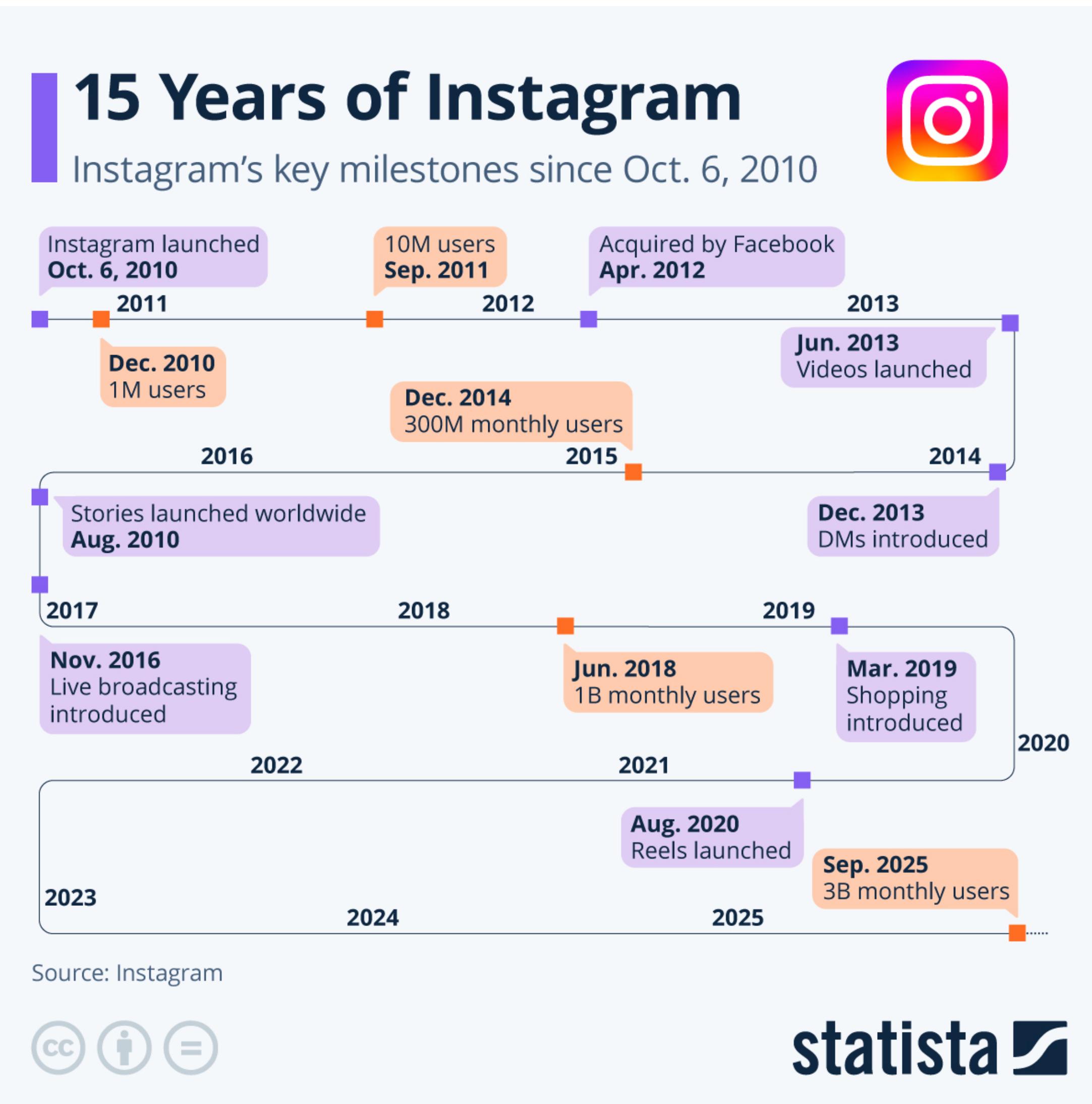
DualMindVLM:

**Long Thinking:** This image depicts a narrow urban street with pedestrians, vehicles, and various signs. Potential hazards include:

1. **Pedestrian-Vehicle Conflict**: The street appears narrow, and vehicles are close to pedestrians. This can lead to accidents if drivers are not cautious.
  - **Mitigation**: Drivers should slow down and be vigilant, especially in areas with pedestrians. Pedestrians should also be cautious and use designated crosswalks.
2. **Limited Visibility**: The narrowness of the street may reduce visibility for both drivers and pedestrians.
  - **Mitigation**: Drivers should maintain a safe distance from the sides of the road and be prepared to stop quickly. Pedestrians should be aware of their surroundings and avoid obstructing the view.
3. **Signage Overload**: There are multiple signs, which might be confusing for drivers and pedestrians.
  - **Mitigation**: Ensure that signs are clear and not overcrowded. Drivers should focus on the road and not be distracted by excessive signage.
4. **Traffic Congestion**: The street seems to have a mix of traffic, which can lead to congestion.
  - **Mitigation**: Implement traffic management strategies such as traffic lights, one-way streets, or designated lanes to manage traffic flow.
5. **Building Obstructions**: The buildings on either side of the street are close, which might affect visibility.
  - **Mitigation**: Ensure that buildings do not obstruct the view of the road, especially for drivers.

By addressing these potential hazards with appropriate mitigations, the safety of pedestrians and drivers can be improved in this urban setting. **(320 tokens)**

# Application — Analyze infographic



Question: When was Instagram acquired by Facebook?

DualMindVLM:

**Short Thinking:** Identify the event labeled "Acquired by Facebook" on the timeline. Final Answer: Apr. 2012 (**25 tokens**)

Question: Summarize Instagram's growth pattern based on the timeline—does the platform show steady, rapid, or irregular expansion?

DualMindVLM:

**Long Thinking:** To summarize Instagram's growth pattern, we need to analyze the timeline for key milestones and user numbers over the years.

- \*\*Initial Launch (2010-2011)\*\*: Instagram launched in October 2010 with 1 million users by December 2010. By September 2011, it reached 10 million users. This indicates a rapid growth in the early years.
- \*\*2012-2013\*\*: The platform was acquired by Facebook in April 2012, which could be seen as a significant milestone. By June 2013, videos were launched, and the user base continued to grow.
- \*\*2014-2015\*\*: In December 2014, it reached 300 million monthly users, showing a steady increase in user base.
- \*\*2016-2019\*\*: The platform introduced Stories in August 2016, live broadcasting in November 2016, and shopping in March 2019. By June 2018, it reached 1 billion monthly users, indicating a continued rapid expansion.
- \*\*2020-2025\*\*: Reels were launched in August 2020, and the goal is to reach 3 billion monthly users by September 2025.

The timeline shows a consistent pattern of rapid growth, with significant milestones leading to substantial increases in user numbers and features. Therefore, the platform shows a rapid expansion pattern.

**Final Answer:** Rapid expansion. (**370 tokens**)

# Application — Analyze stock chart

## Stocks Emerge From Covid Crash With Historic 12-Month Run

Performance of major U.S. stock market indices since January 2020 (indexed to closing prices on March 23, 2021)



Source: Yahoo! Finance



statista

Question: Which color represents the S&P 500 index?

DualMindVLM:

**Short Thinking:** The legend at the top of the graph indicates the color for each index. The S&P 500 is represented by the orange color.

**Final Answer:** Orange (**36 tokens**)

Question: By examining the spacing between the three curves, what can be inferred about sector performance differences, especially regarding technology-heavy vs traditional industries?

DualMindVLM:

**Long Thinking:** To analyze the performance differences between technology-heavy and traditional industries, we need to compare the trends of the Nasdaq (representing technology-heavy sectors) with the S&P 500 and Dow Jones (representing more traditional industries).

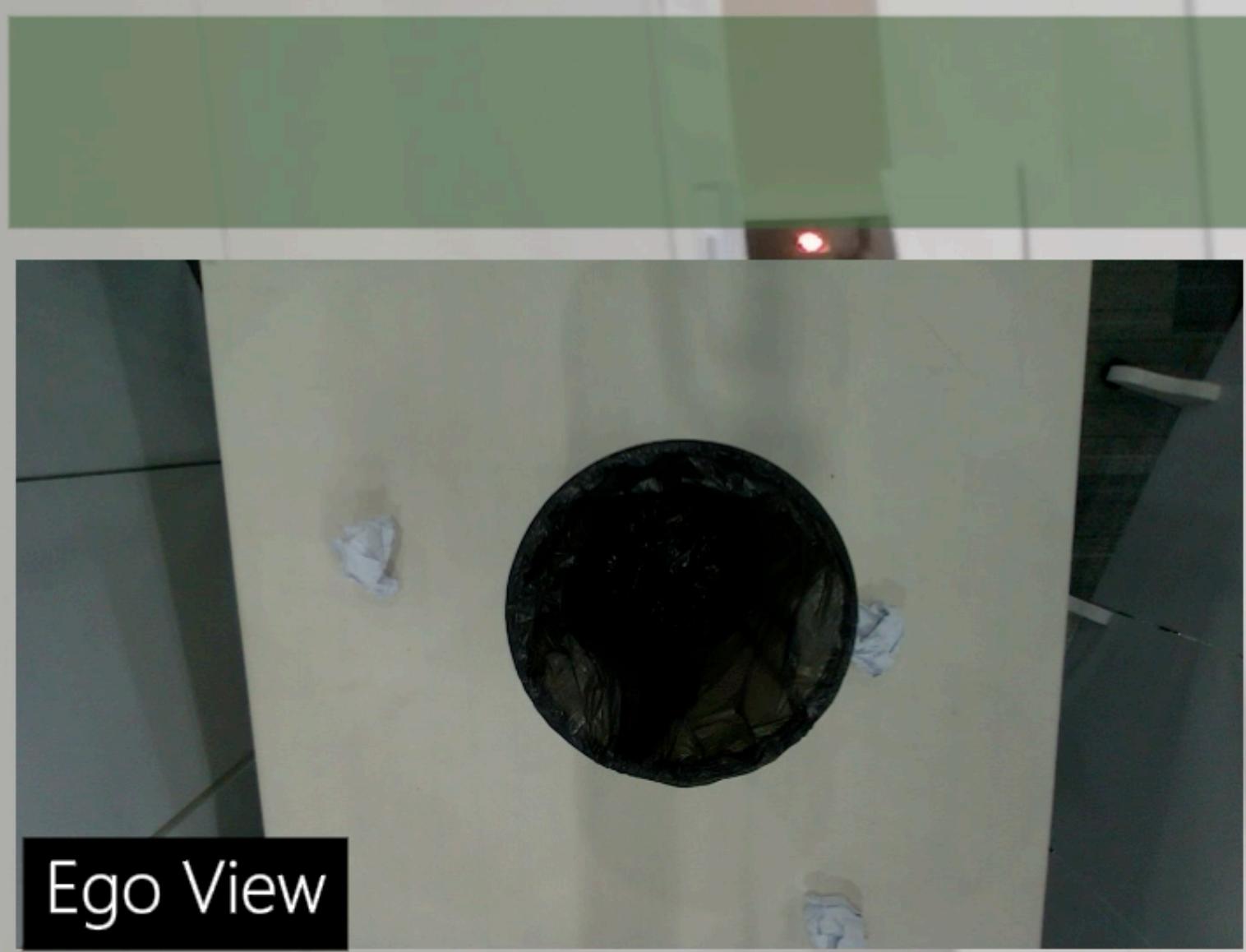
1. **Trend Analysis:** The Nasdaq curve shows a steeper and more consistent upward trend compared to the S&P 500 and Dow Jones. This suggests that technology-heavy sectors have performed better than traditional industries over the period.
2. **Magnitude of Growth:** The Nasdaq has the highest percentage increase (+94.99%), indicating a stronger recovery and growth in technology sectors. The S&P 500 and Dow Jones have similar growth rates (+76.12% and +76.05%, respectively), but they are still lower than the Nasdaq.
3. **Volatility:** The Nasdaq curve shows more volatility, which is typical for technology stocks that are more sensitive to market changes and innovations. However, this volatility is reflected in the overall upward trend, indicating resilience and growth.

**Final Answer:** The spacing between the curves indicates that technology-heavy sectors, represented by the Nasdaq, have outperformed traditional industries, as shown by the higher percentage increase and the more consistent upward trend. (**261 tokens**)

## Application — AI Assistant



## Application — Embodied AI

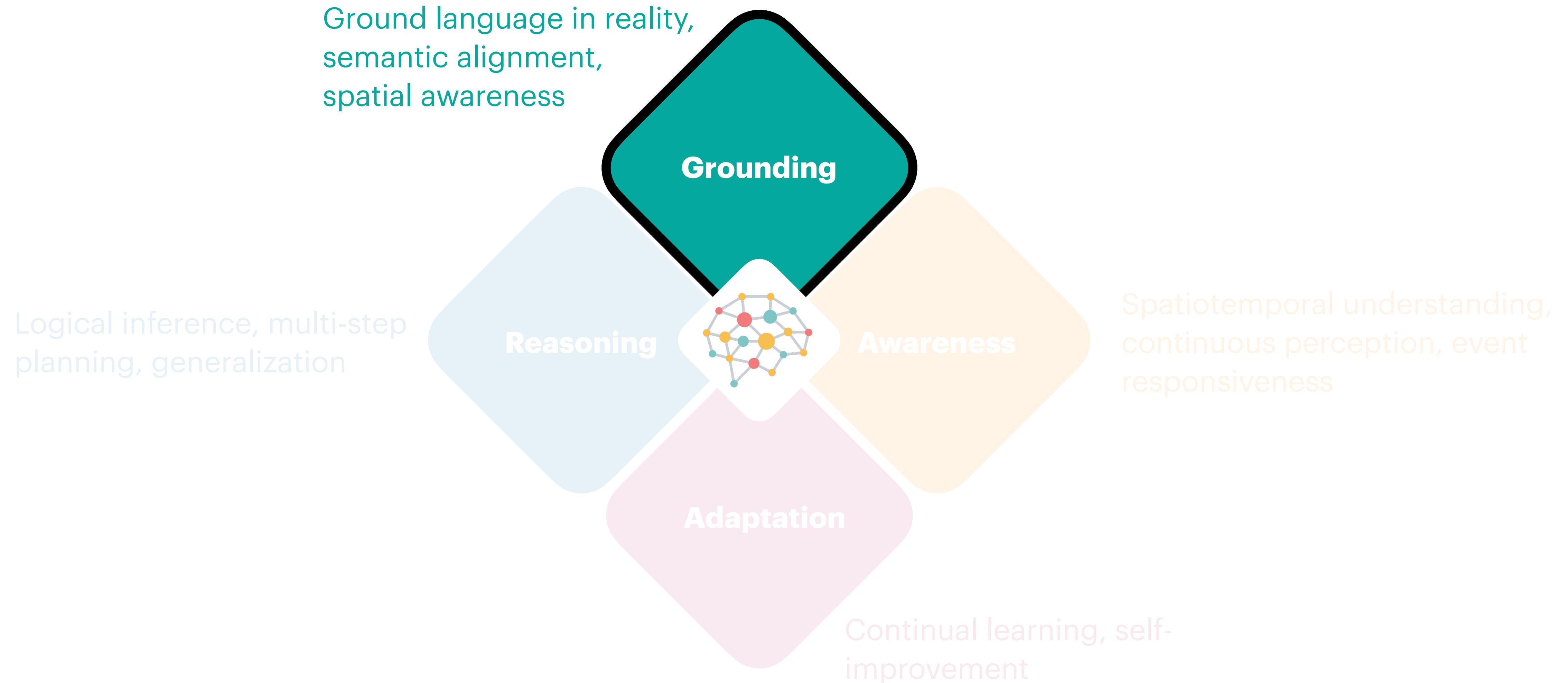


Ego View



w/o fine-tuning on robotic data!!

# Cornerstones of multimodal agents

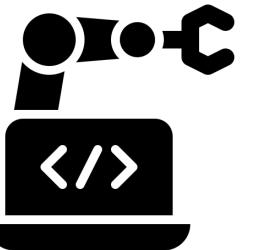


# Robotic Manipulation

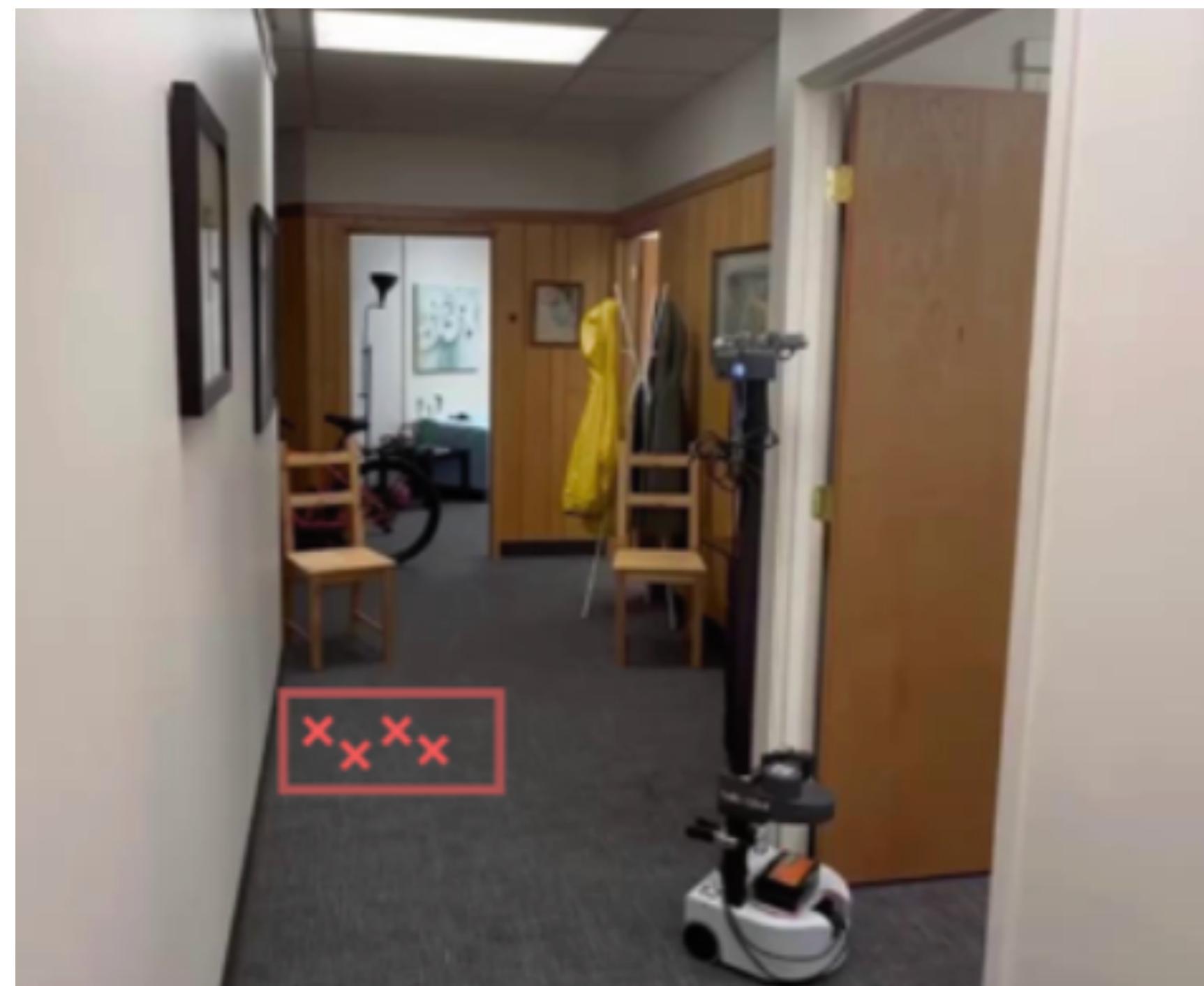


User:  
Place the LEGO block  
**within the left part of the**  
pizza box.

Agent:  
[(0.35, 0.88), (0.43, 0.88),  
(0.38, 0.89), (0.45, 0.87),  
(0.47, 0.88)].

  
`</>`  
`pick()`  
`place()`  
`push()`  
`go_to()`  
`follow_path()`  
`walk_to()`  
...

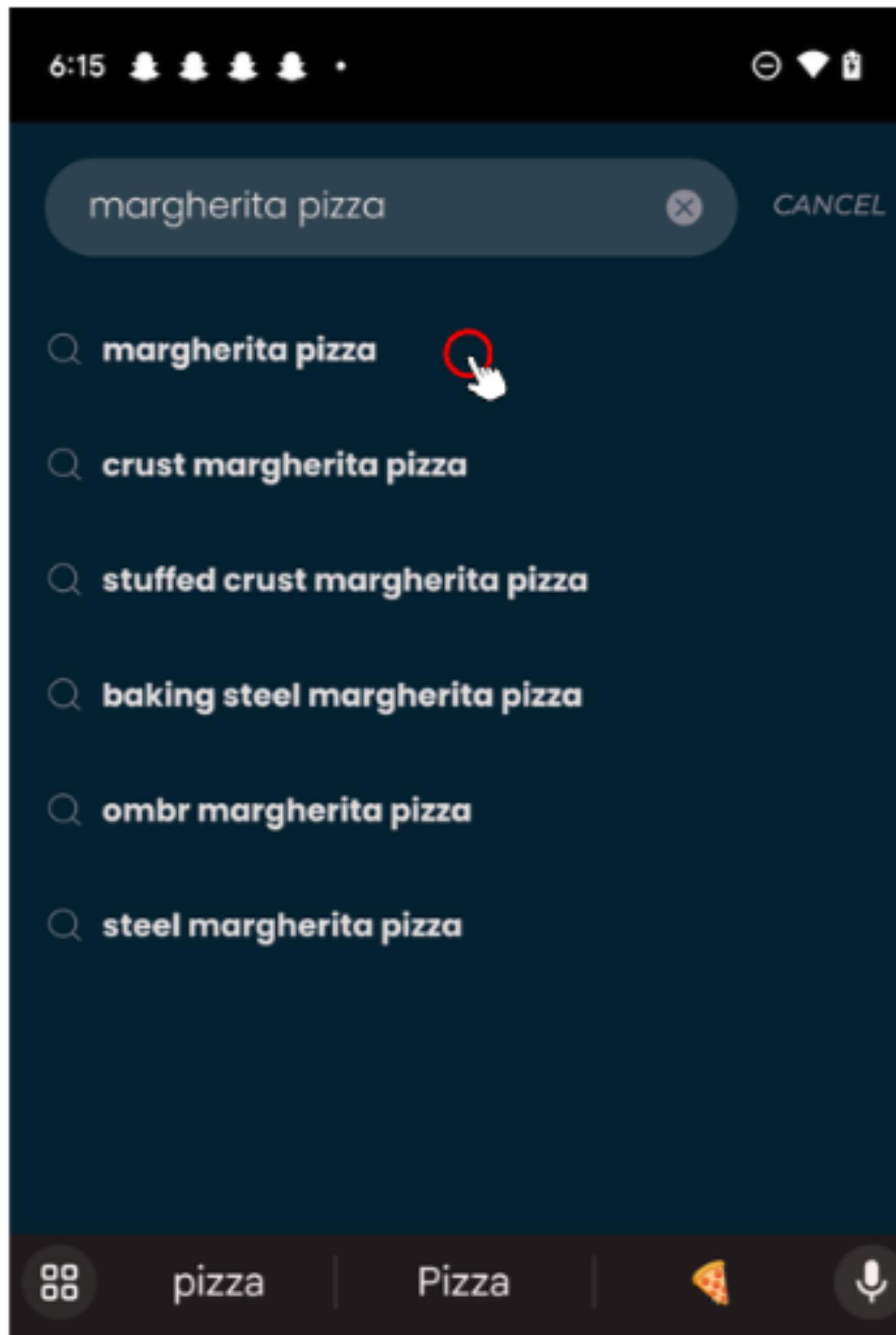
# Robotic Navigation



User:  
Navigate to the empty  
space **in front of** the  
chair.

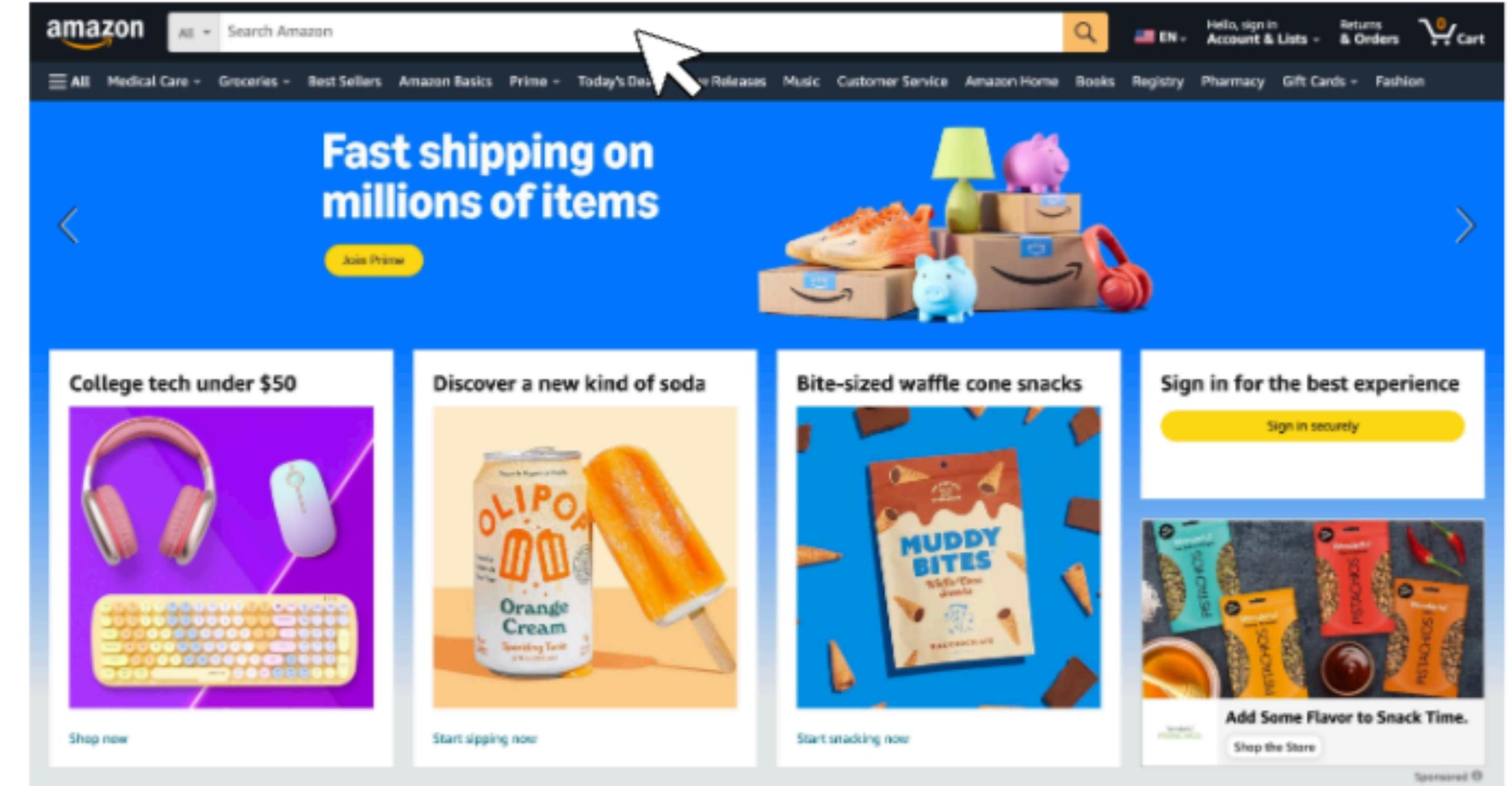
Agent:  
[(0.32, 0.64), (0.34, 0.62),  
(0.34, 0.68), (0.30, 0.65),  
(0.32, 0.69)].

# GUI Agent



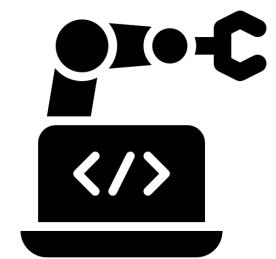
"element": "the first search result labeled 'margherita pizza'"

"click": (540, 399)

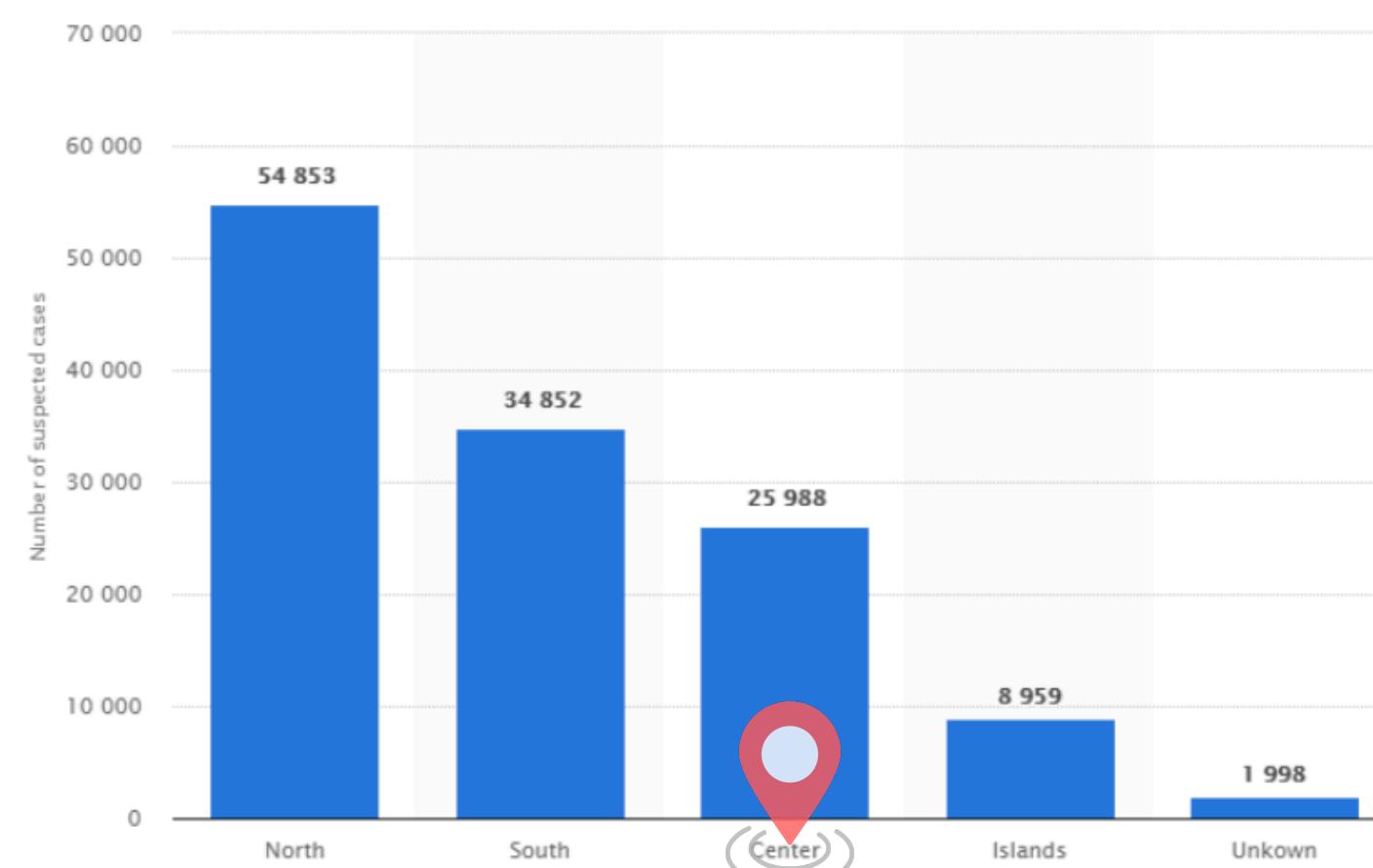


"element": "the **search bar** at the top of the page"

"click": (556, 26)



moveTo(), dragTo(), click(), scroll(), ...



square beads	\$2.97 per kilogram
oval beads	\$3.41 per kilogram
flower-shaped beads	\$1.83 per kilogram
star-shaped beads	\$1.95 per kilogram
heart-shaped beads	\$1.52 per kilogram
spherical beads	\$3.42 per kilogram
rectangular beads	\$1.97 per kilogram



Grounding improves accuracy & explainability

Revenue from external customers, classified by significant product and service offerings, was as follows:

(in millions)

Year Ended June 30,	2019	2018	2017
Server products and cloud services	32,622	26,129	21,649
Office products and cloud services	31,769	28,316	25,573
Windows	20,395	19,518	18,593
Gaming	11,386	10,353	9,051
Search advertising	7,628	7,012	6,219
<b>LinkedIn</b>	<b>6,754</b>	<b>5,259</b>	<b>2,271</b>
<b>Enterprise Services</b>	<b>6,104</b>	<b>5,846</b>	<b>5,542</b>
<b>Devices</b>	<b>6,095</b>	<b>5,134</b>	<b>5,062</b>
<b>Other</b>	<b>3,070</b>	<b>2,793</b>	<b>2,611</b>
Total	<b>\$125,843</b>	<b>\$110,360</b>	<b>\$96,571</b>

Our commercial cloud revenue, which includes Office 365 Commercial, Azure, the commercial portion of LinkedIn, Dynamics 365, and other commercial cloud properties, was **\$38.1 billion**, **\$26.6 billion** and **\$16.2 billion** in fiscal years 2019, 2018, and 2017, respectively. These amounts are primarily included in Office products and cloud services, Server products and cloud services, and LinkedIn in the table above.

PERNIAGAAN ZHENG HUI

JM0328955-V  
NO.59 JALAN PERMAS 9/5  
BANDAR BARU PERMAS JAYA  
81750 JOHOR BAHRU  
TEL : 07-386 7524 FAX : 07-386 3793  
GST NO : 000800589824

#### SIMPLIFIED TAX INVOICE

GOGIANT ENGINEERING (M) SDN BHD			
Receipt #:	CS00082258	Date:	09/02/2018
Salesperson:		Time:	08:32:00
Cashier:	USER		
Item	Qty	(RM) RSP	(RM) Amount
6783	1809	5 3.50	17.50
SR: CERAMIC CAP			
2954	1809	30 3.80	114.00
SR: S/STEEL 1/2" STREET ELBOW			
1760	6A	1 55.00	55.00
SR: 2.4MM STARWELD RED HEAD TUNGSTEN ROD			
3496	6A	3 33.00	99.00
SR: ESICUT 4" CUTTING DISC (1BOX)-50PCS			
2460	6A	2 43.00	86.00
SR: WELDRO PICKLING GEL 1KG			
9428	6B	4 10.00	40.00
SR: 13.5" WELDING GLOVE - GREEN (GS)			
TOT QTY: 45			411.50
(Excluded GST) Sub Total (RM) :			411.50
Discount (RM) :			0.00
Total GST (RM) :			24.69
Rounding (RM) :			0.01
Total (RM) :			436.20
CASH :			436.20
Change (RM) :			0.00

#### GST SUMMARY

Tax Code	%	Amount	GST
SR	6	411.50	24.69
Total :		411.50	24.69

GOODS SOLD ARE NOT RETURNABLE,  
THANK YOU.

# Grounding could go wrong

tomato sauce	\$2.47 per kg
beef sauce	\$1.85 per kg
marinara sauce	\$1.17 per kg
alfredo sauce	\$1.62 per kg

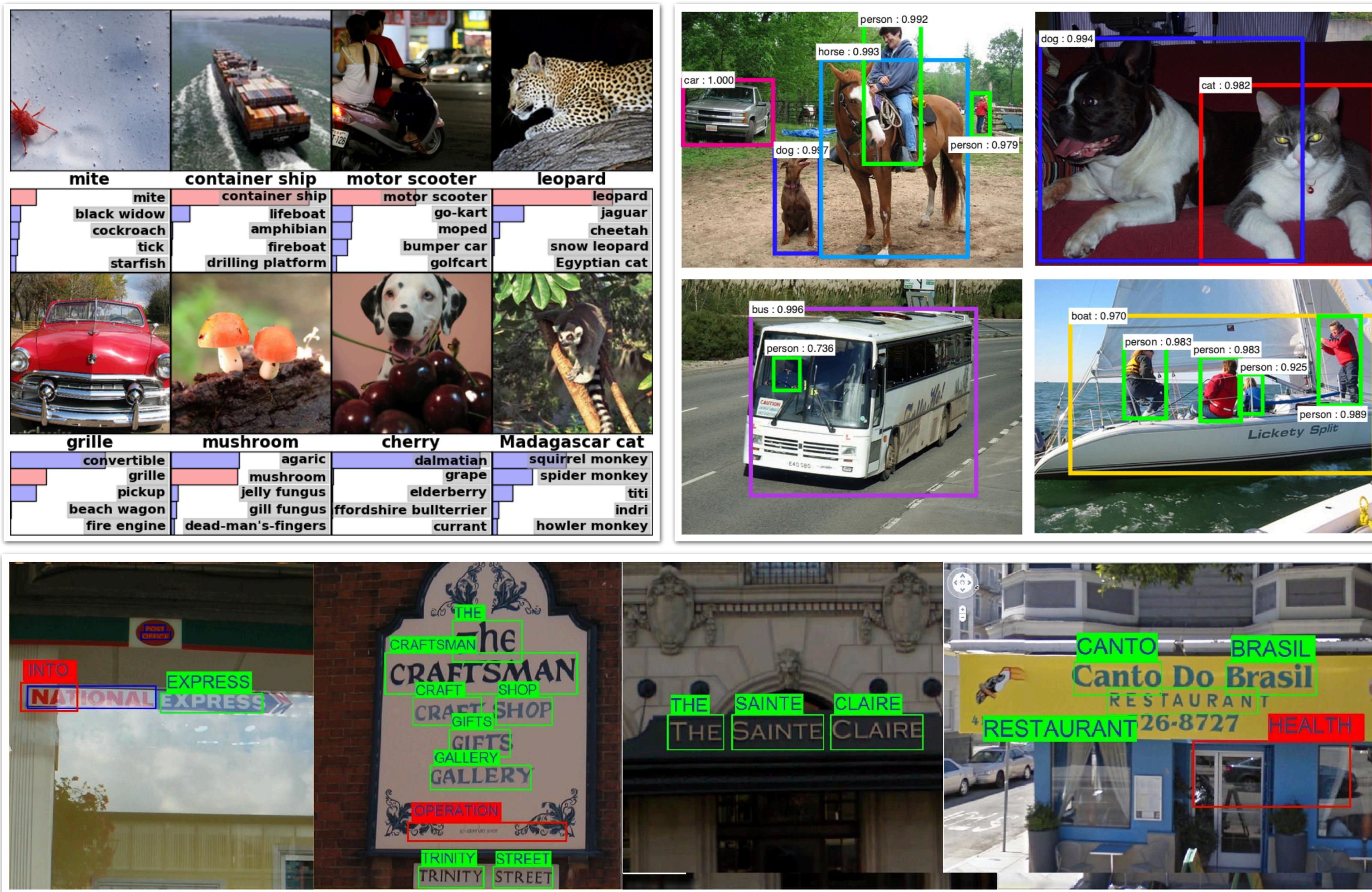
User:

Felix wants to buy 2 kilograms of beef sauce and 2 kilograms of marinara sauce. How much will he spend? *Answer: 6.04*

AI:

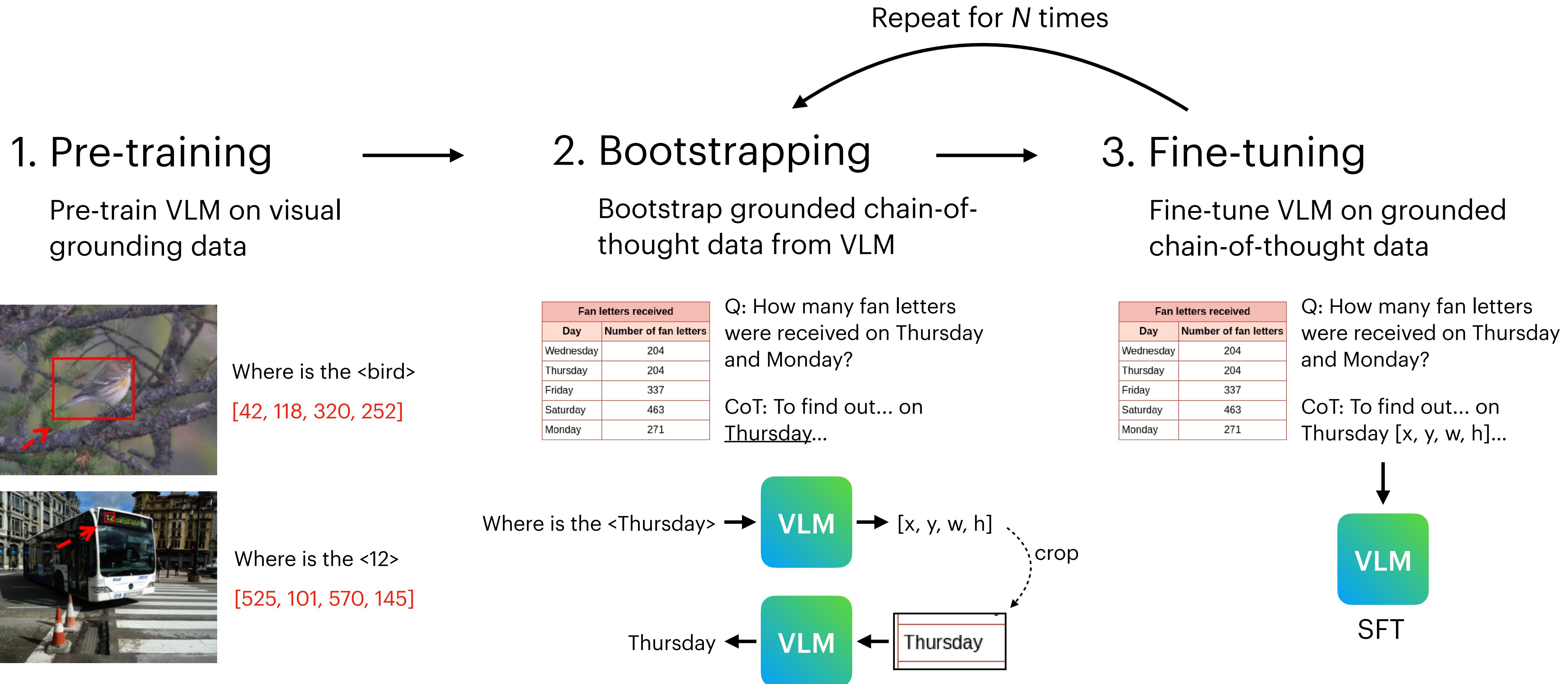
To find out how much Felix will spend on 2 kilograms of **beef sauce** [0.231, 0.485, 0.445, 0.575], we need to multiply the price per kilogram by the number of kilograms he is buying. The price per kilogram of beef sauce is \$1.17 [0.631, 0.492, 0.787, 0.577]. The total cost of 2 kilograms of beef sauce would be:  $\$1.17 * 2 = \$2.34$ . To find out how much Felix will spend on 2 kilograms of **marinara** [0.011, 0.531, 0.331, 0.595] sauce, we need to multiply the price per kilogram by the number of kilograms he is buying. The price per kilogram of marinara sauce is \$1.62 [0.791, 0.265, 0.975, 0.319]. The total cost of 2 kilograms of marinara sauce would be:  $\$1.62 * 2 = \$3.24$ . Total cost =  $\$2.34 + \$3.24 = \$5.58$ . \*Answer\*: 5.58. 

# Key idea: use **recognition** to verify grounding



[Source: Krizhevsky et al., 2012; Ren et al., 2015; Wang et al., 2012]

# GCoT: Grounded Chain-of-Thought

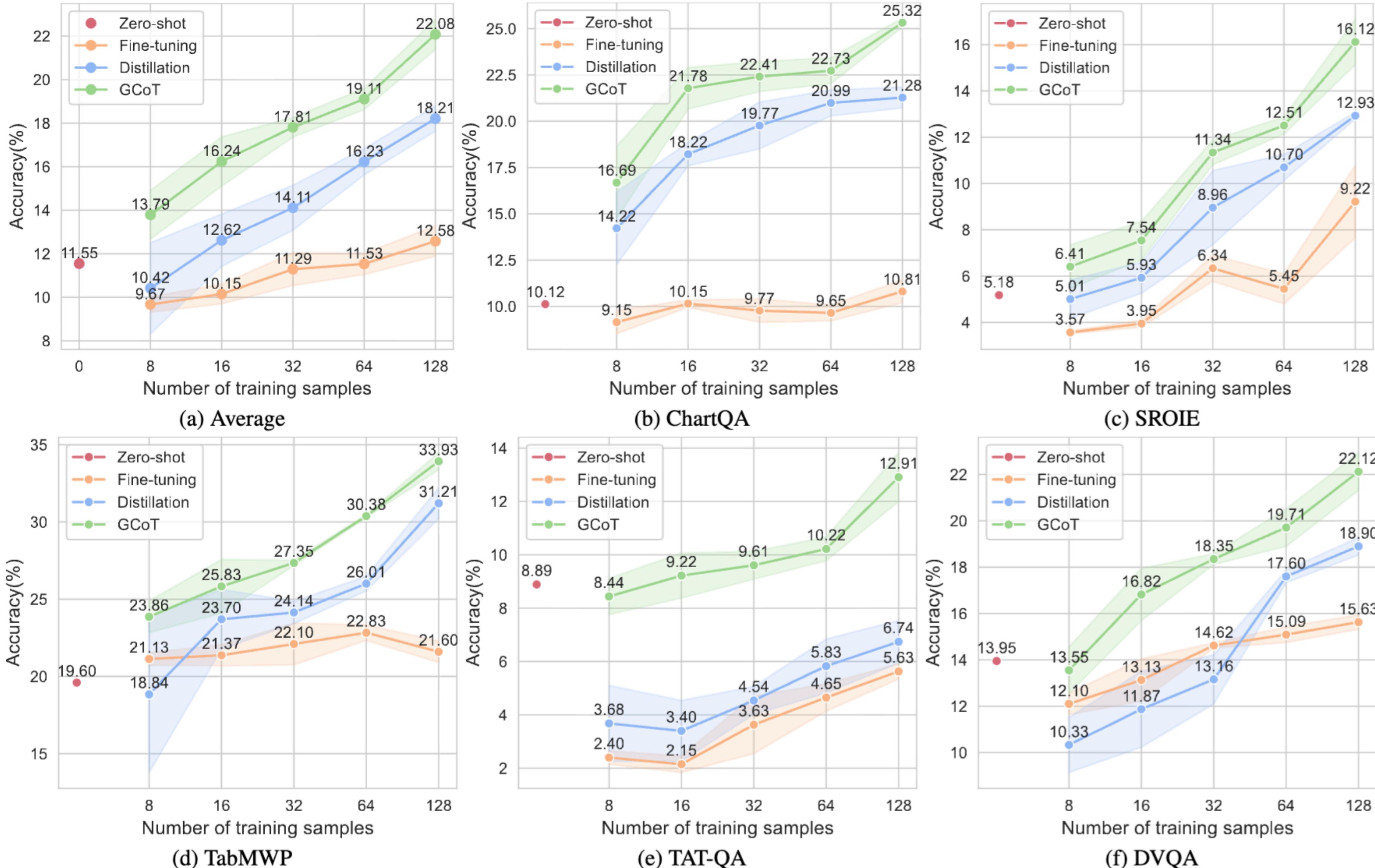


# GCoT in few-shot learning

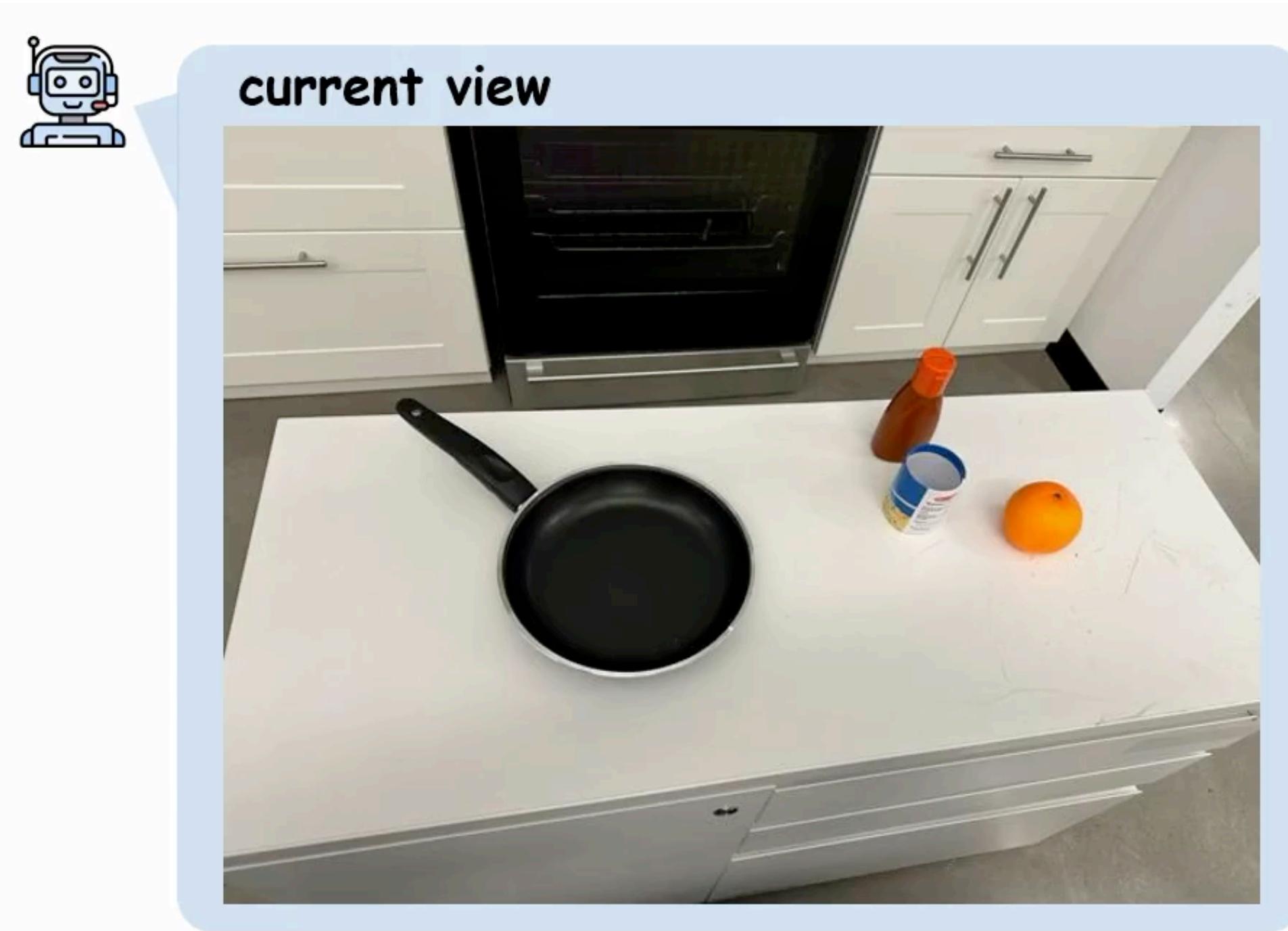
The benchmark covers charts, receipts, tables, and reports.

GCoT significantly beats CoT (distilled), fine-tuning, and zero-shot in few-shot settings.

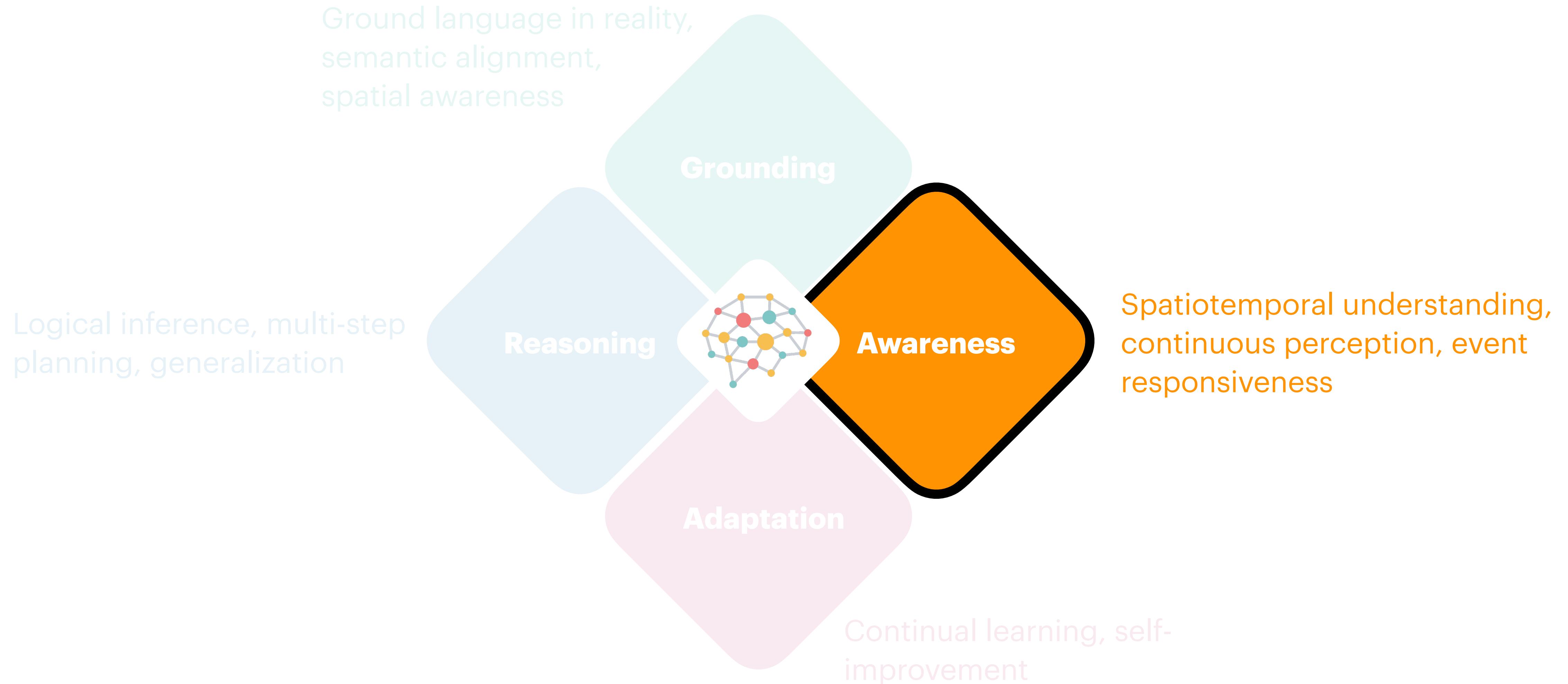
GCoT maintains a healthy uptrend when more data is available.



# GCoT in robot grounding

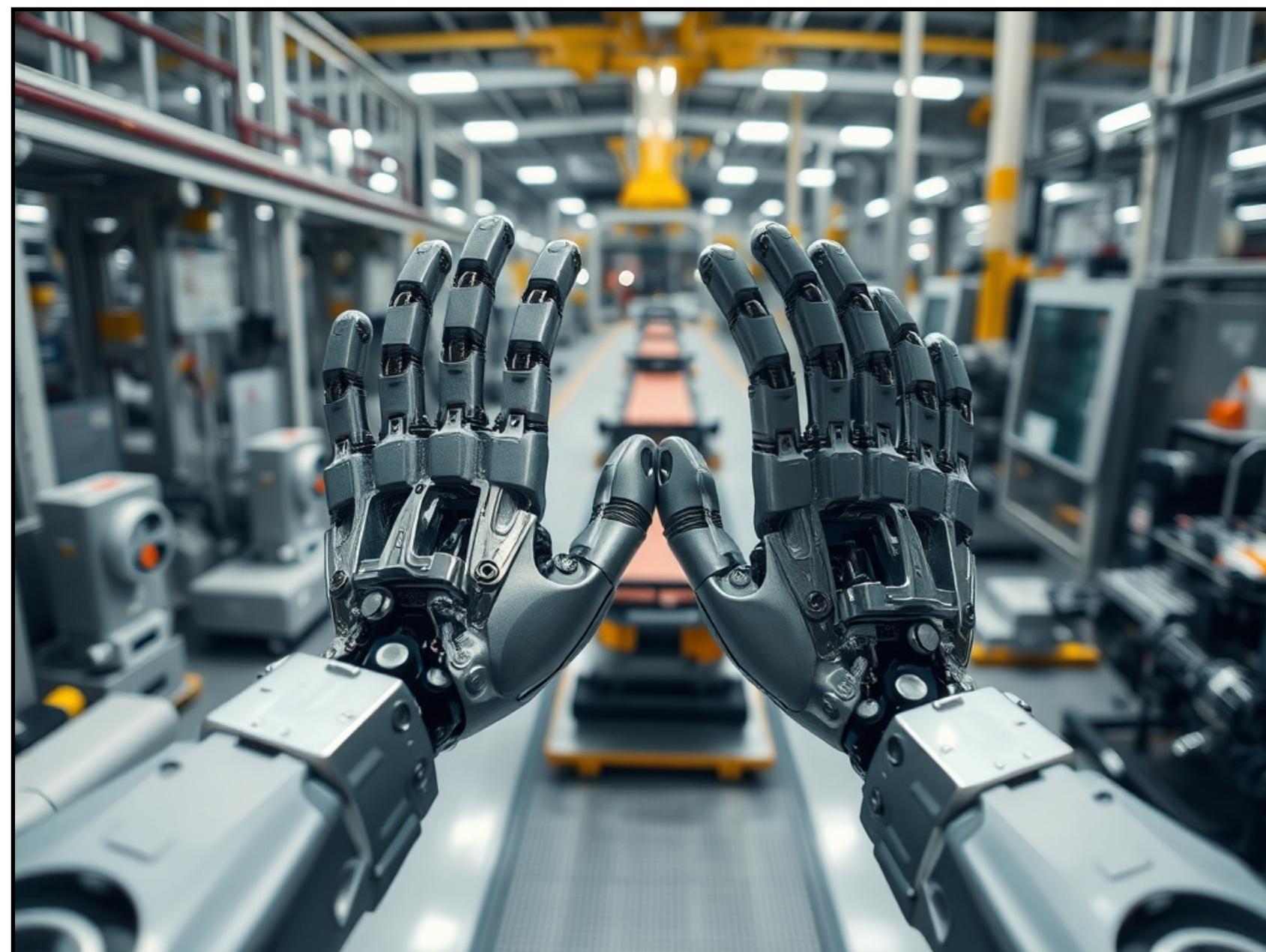


# Cornerstones of multimodal agents



# The real world is inherently streaming

Robotics



Self-driving



AI glasses

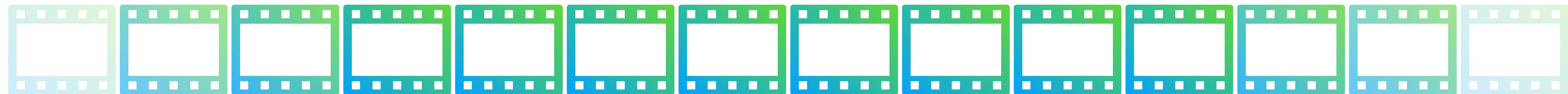


Challenges:

Unbounded data, volume, real-time analysis

Continuous processing, latency issue, dynamic contexts

Variable response timing, immediate/delayed

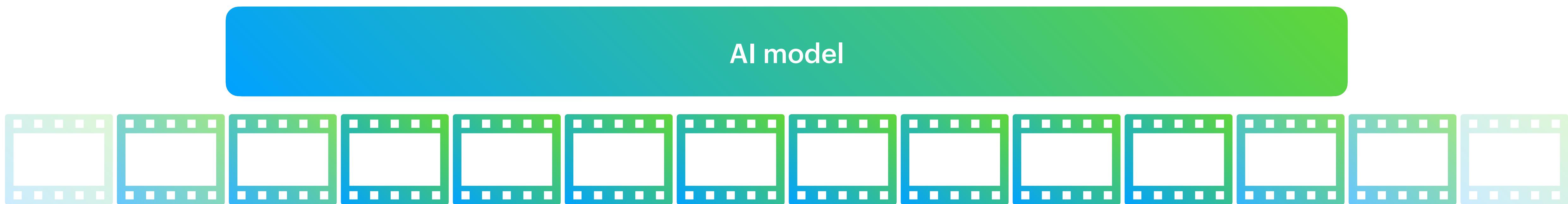


## Challenges:

Unbounded data, volume, real-time analysis

Continuous processing, latency issue, dynamic contexts

Variable response timing, immediate/delayed

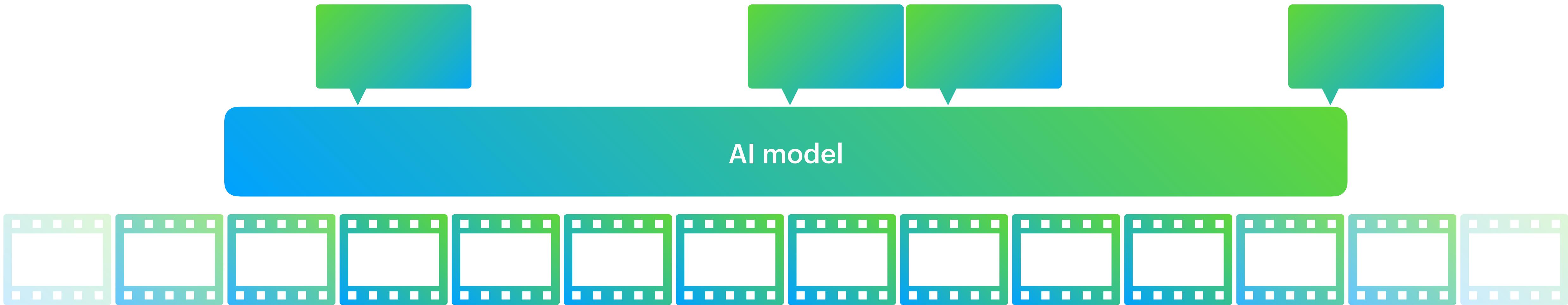


## Challenges:

Unbounded data, volume, real-time analysis

Continuous processing, latency issue, dynamic contexts

Variable response timing, immediate/delayed



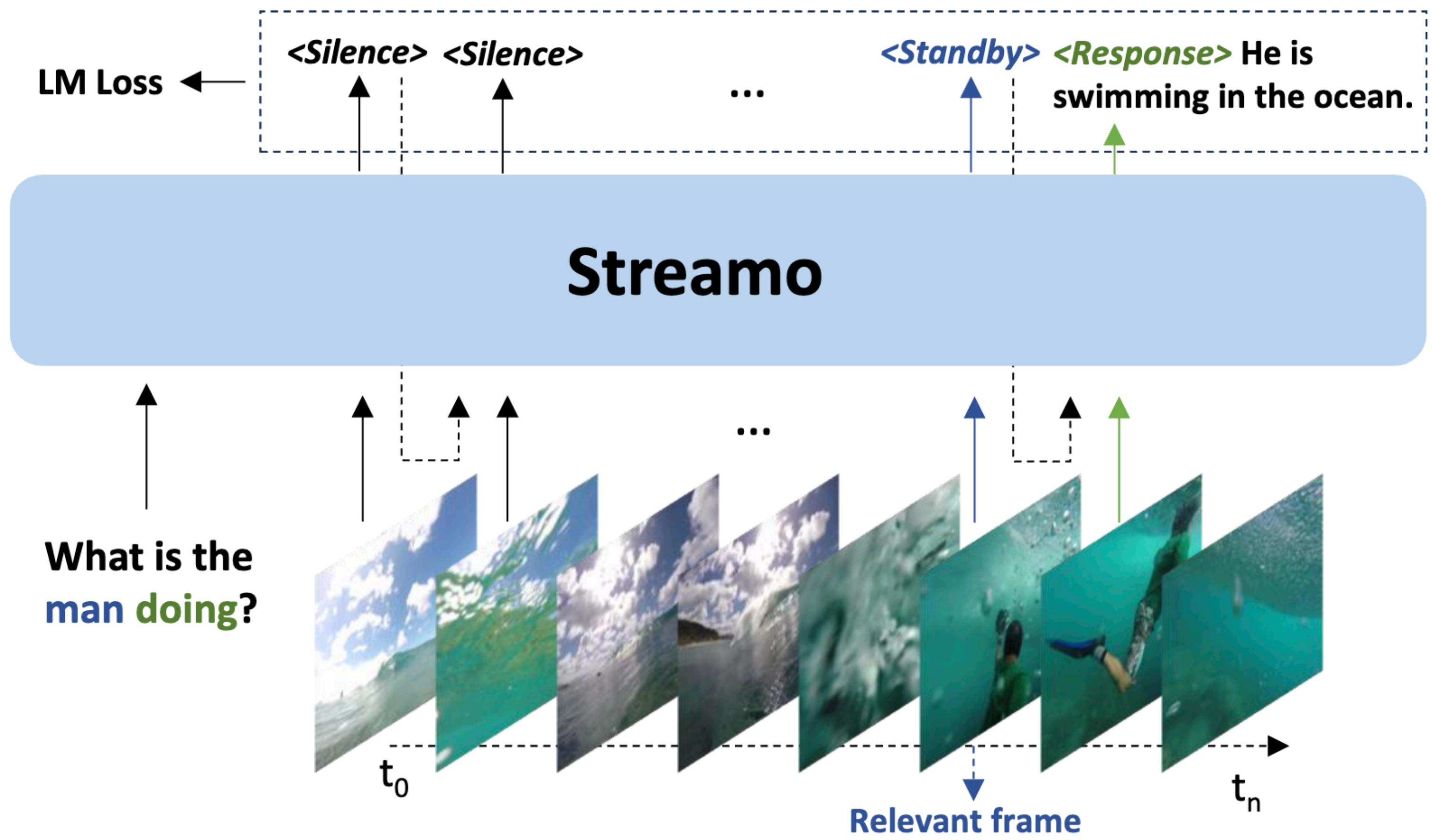
# Streamo: end-to-end, multi-task, multi-modal

**Unifies** perception, decision-making, and response generation in a single model via a data-driven approach.

<Silence>  
remains silence while processing incoming frames.

<Standby>  
detects relevant input and waits for complete information.

<Response>  
generates a response when receiving enough information.

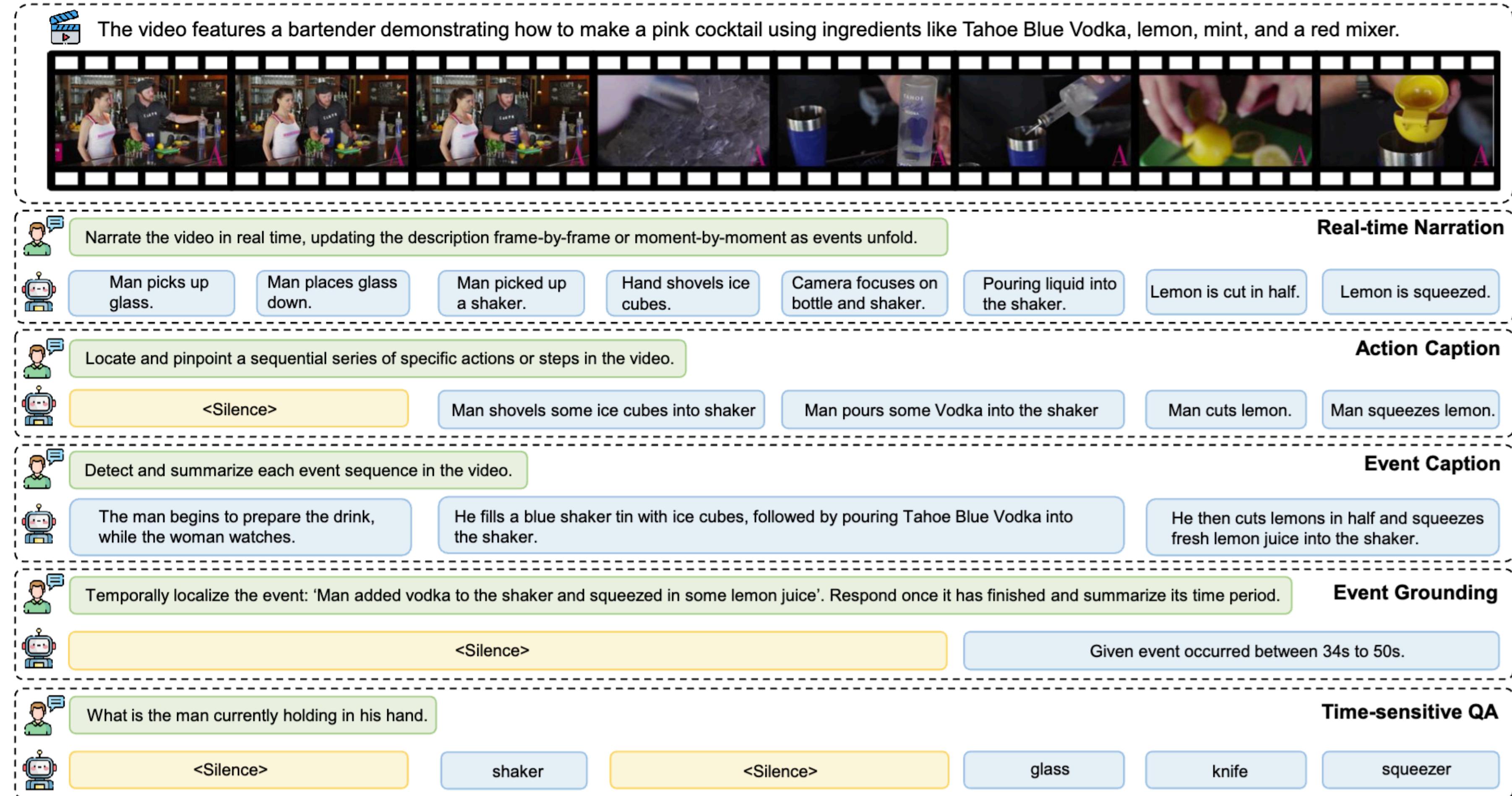


# Streamo-Instruct-465K

A **large-scale, multi-task** streaming video instruction tuning dataset.

**Unified temporal annotation** across tasks, with clear event boundaries and time alignment.

Any offline video model can be transformed into an online model with this dataset.



# Streamo-Instruct-465K

**Real-time narration**

Action caption

Event caption

Event grounding

Time-sensitive QA



# Streamo-Instruct-465K

Real-time narration

**Action caption**

Event caption

Event grounding

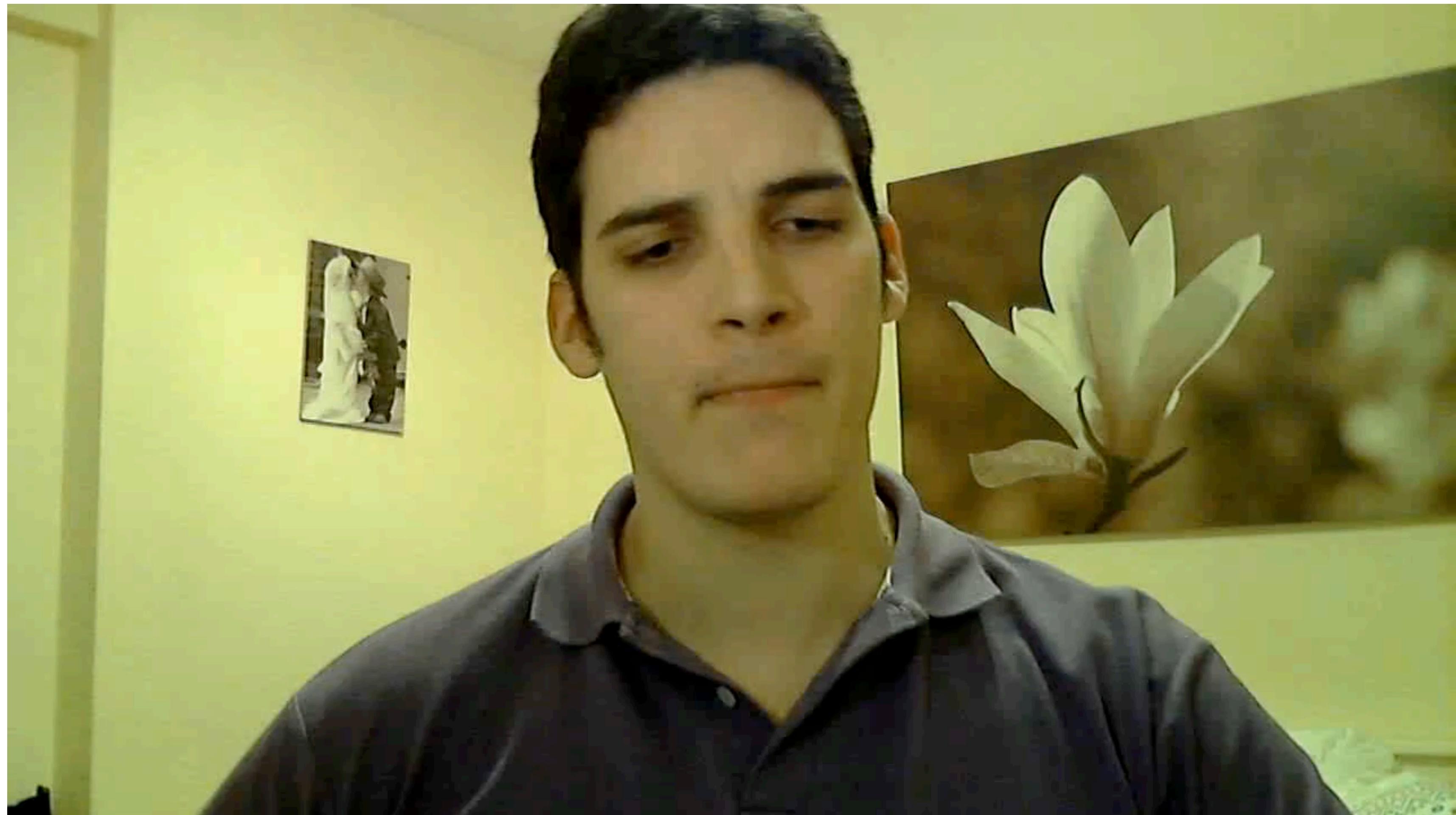
Time-sensitive QA



Prompt: Identify and mark the video segments corresponding to a series of actions or steps.

# Streamo-Instruct-465K

Real-time narration  
Action caption  
**Event caption**  
Event grounding  
Time-sensitive QA



Prompt: Extract and explain all notable activities in the video.

# Streamo-Instruct-465K

Real-time narration

Action caption

Event caption

**Event grounding**

Time-sensitive QA



Prompt: Scan the video for all event of 'the camera zooms in on the waterfall' and provide their precise locations.

# Streamo-Instruct-465K

Real-time narration

Action caption

Event caption

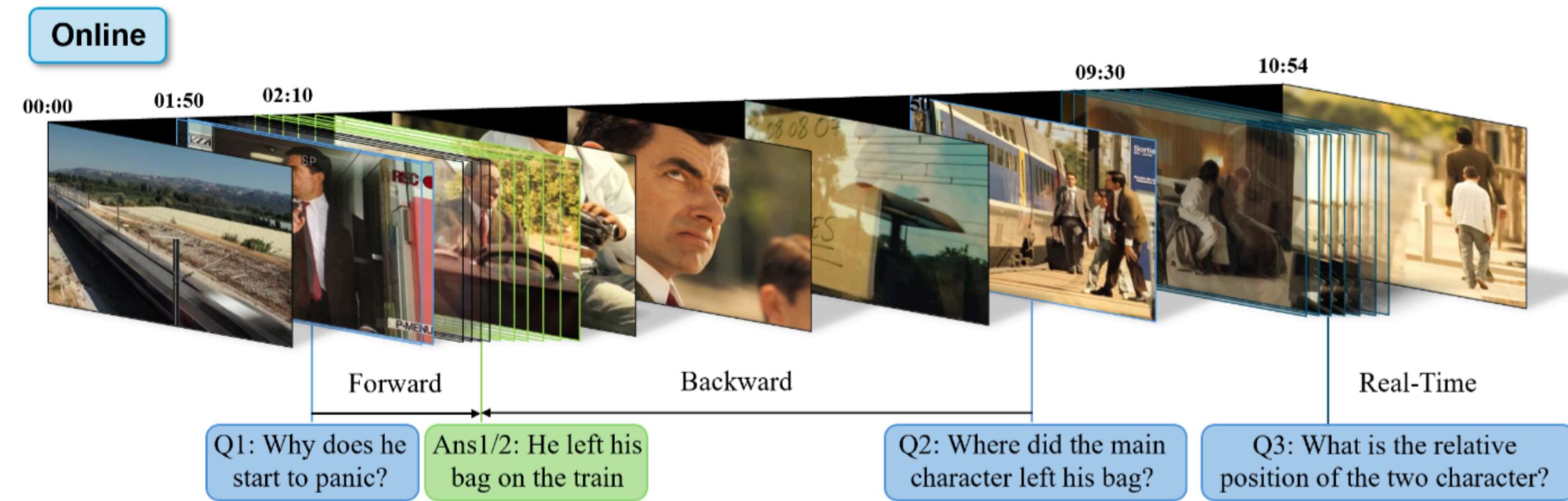
Event grounding

**Time-sensitive QA**



Prompt: What ingredient is being added to the sandwich?

# Quantitative results



[Source: OVO-Bench]

		# frames	realtime perception	backward tracing	forward responding
Offline models	LLaVA-Video	64	63.52	40.40	54.82
	LLaVA-OneVision	64	64.02	43.71	50.50
	Qwen-VL	64	55.98	46.46	48.74
	InternVL-V2	64	60.39	43.44	46.60
	LongVU	1fps	57.61	35.01	47.50
Online models	Flash-VStream	1fps	28.37	27.38	45.09
	VideoLLM-online	2fps	20.79	17.73	N/A
	Dispider	1fps	54.55	36.06	34.72
	<b>Streamo</b>	2fps	<b>67.44</b>	<b>49.18</b>	<b>56.96</b>

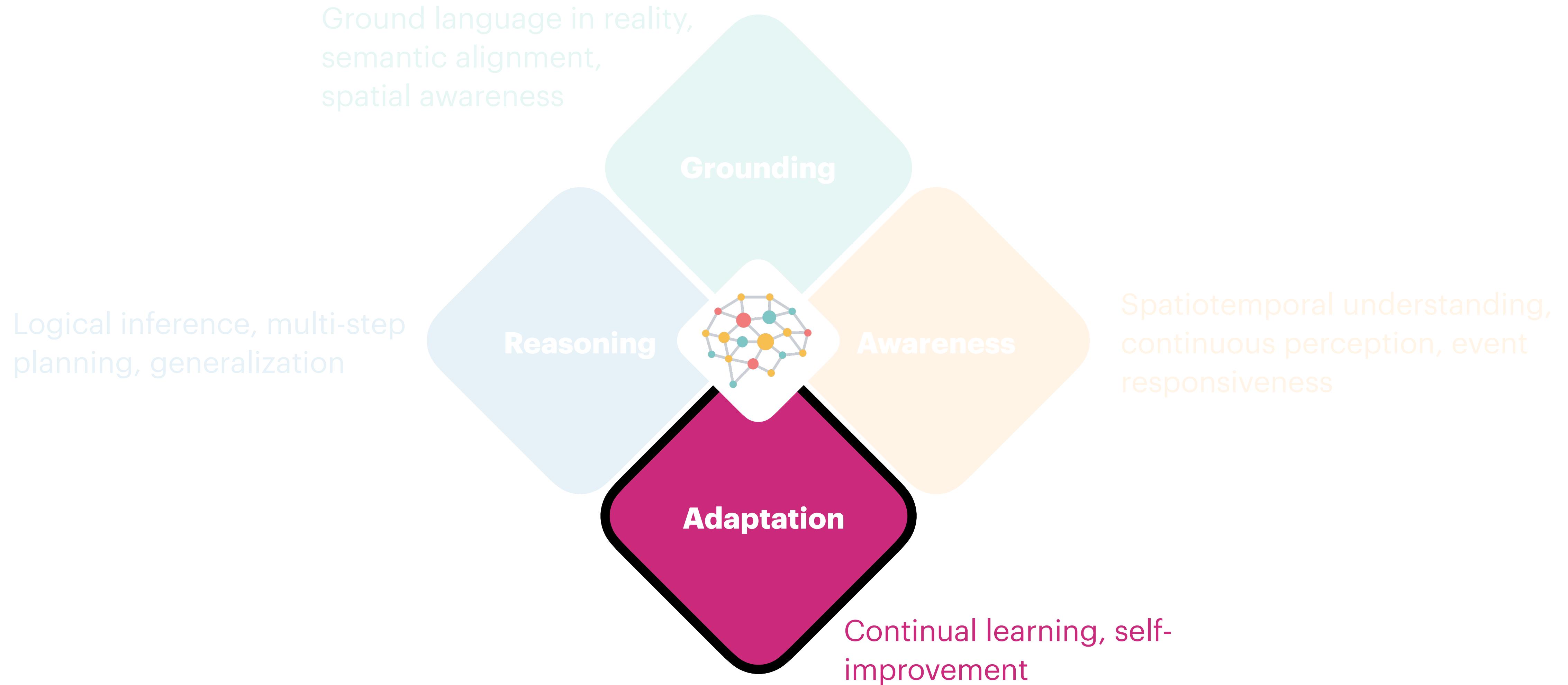
>> 2X Speed

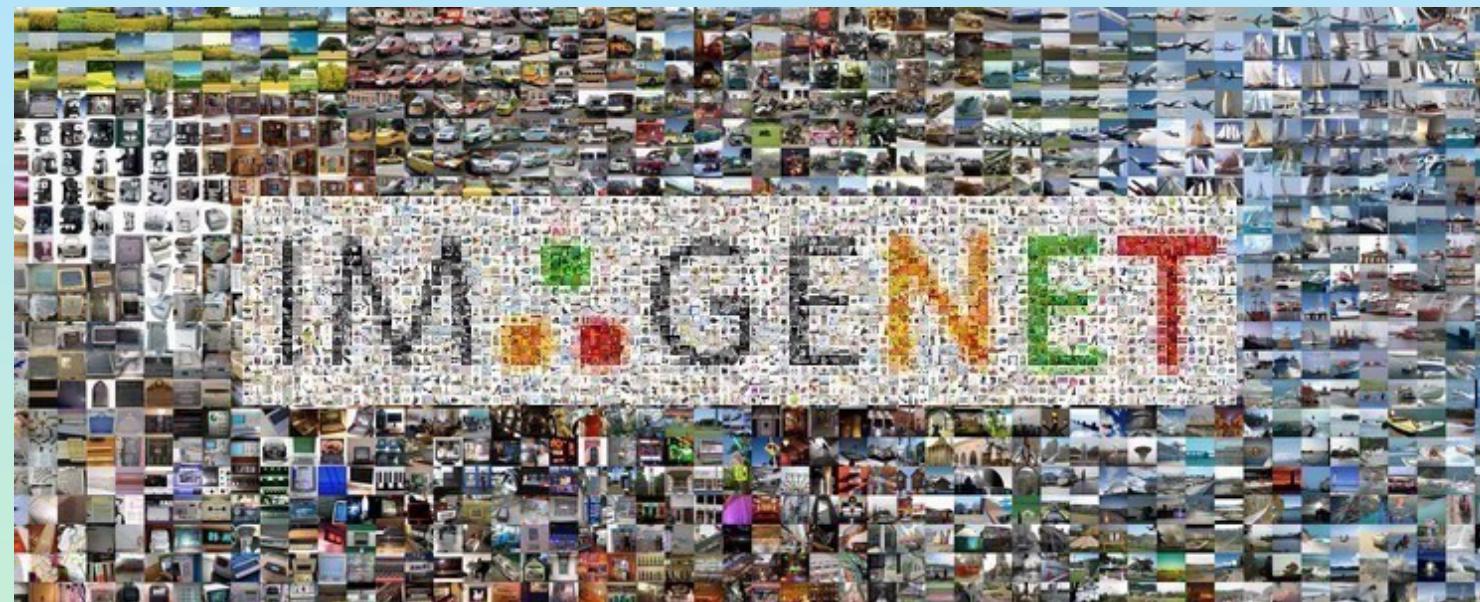


# *Titli's Busy Kitchen*

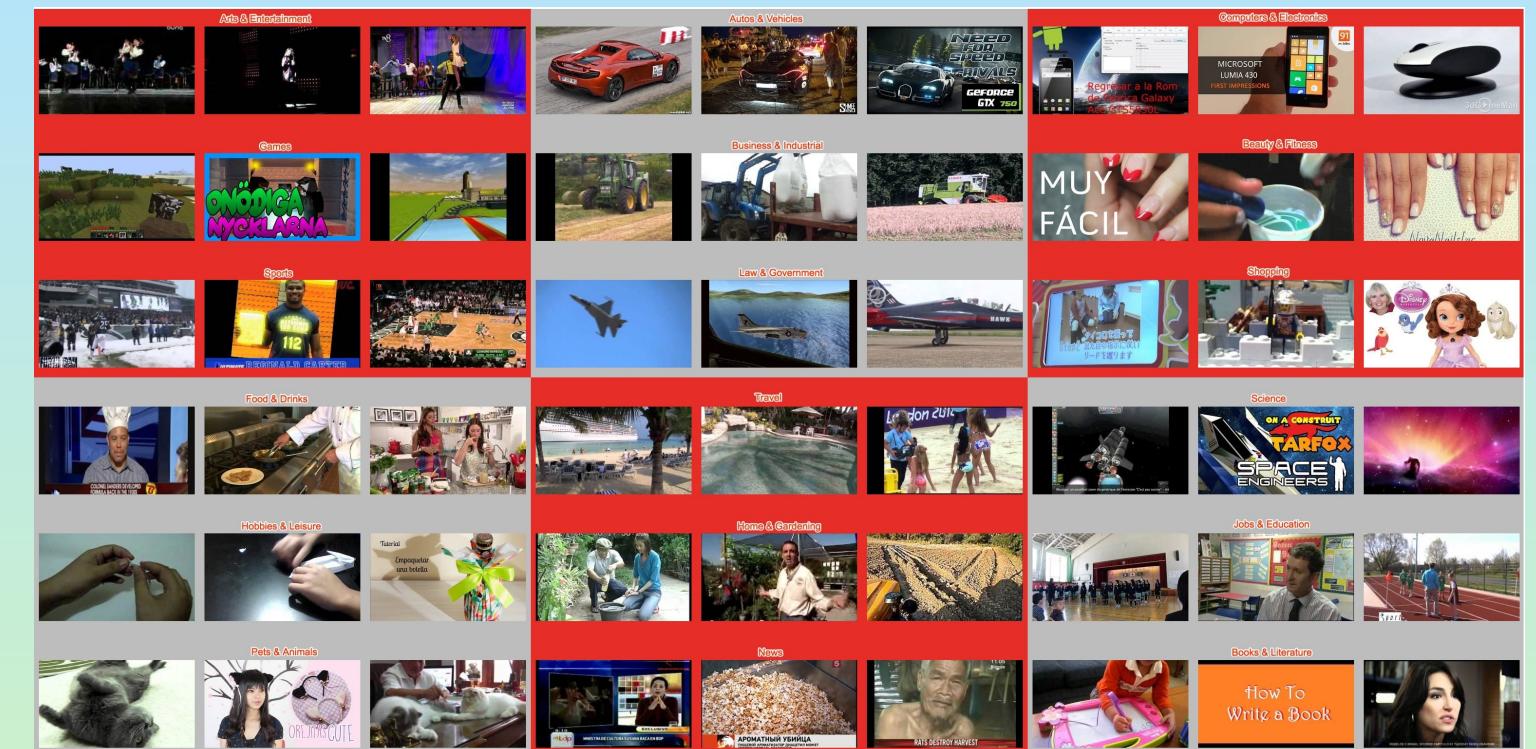


# Cornerstones of multimodal agents





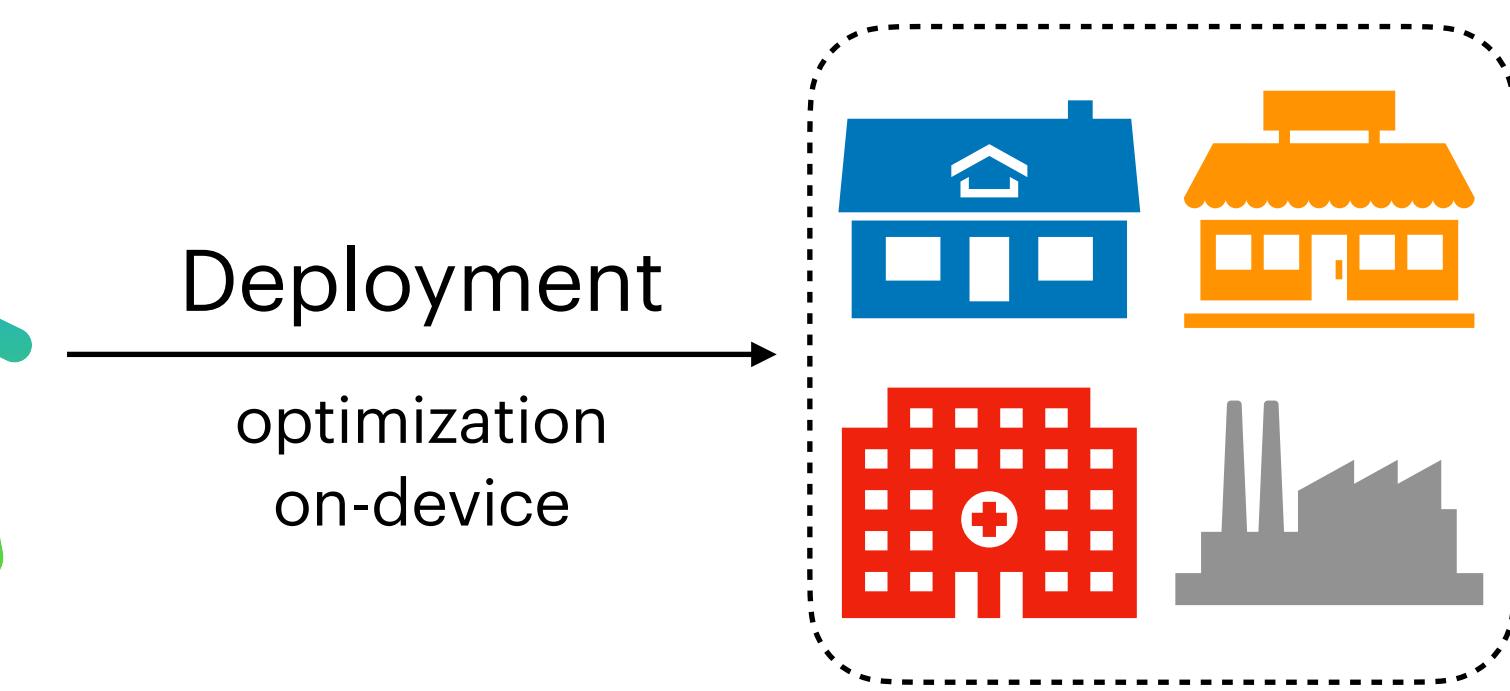
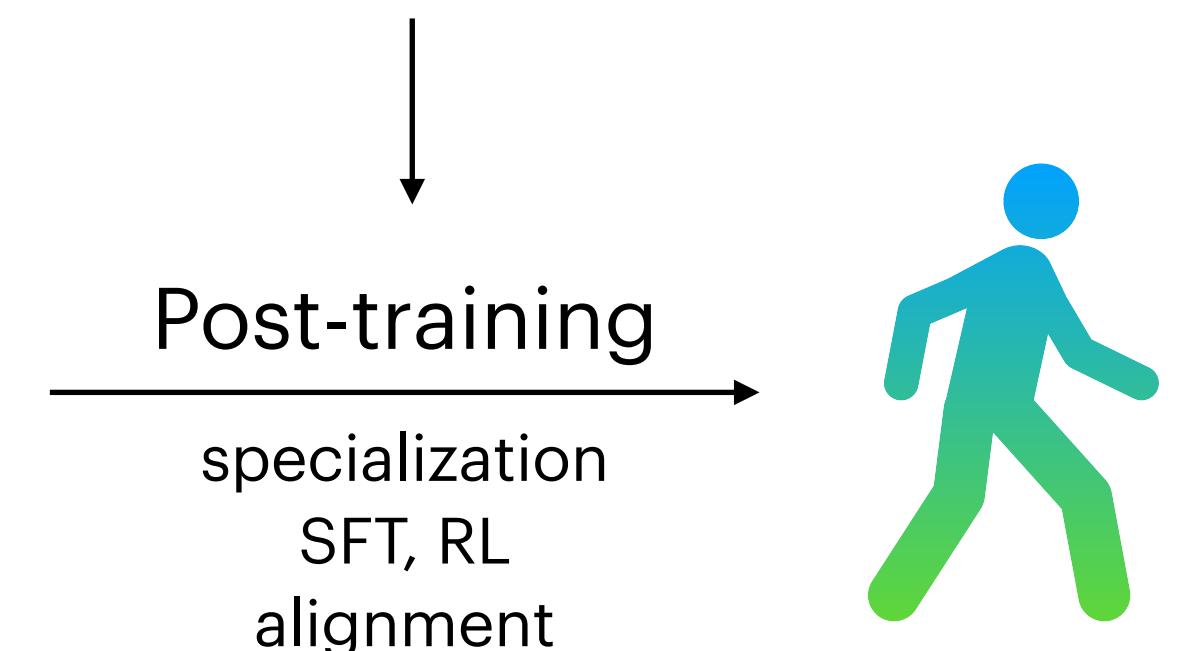
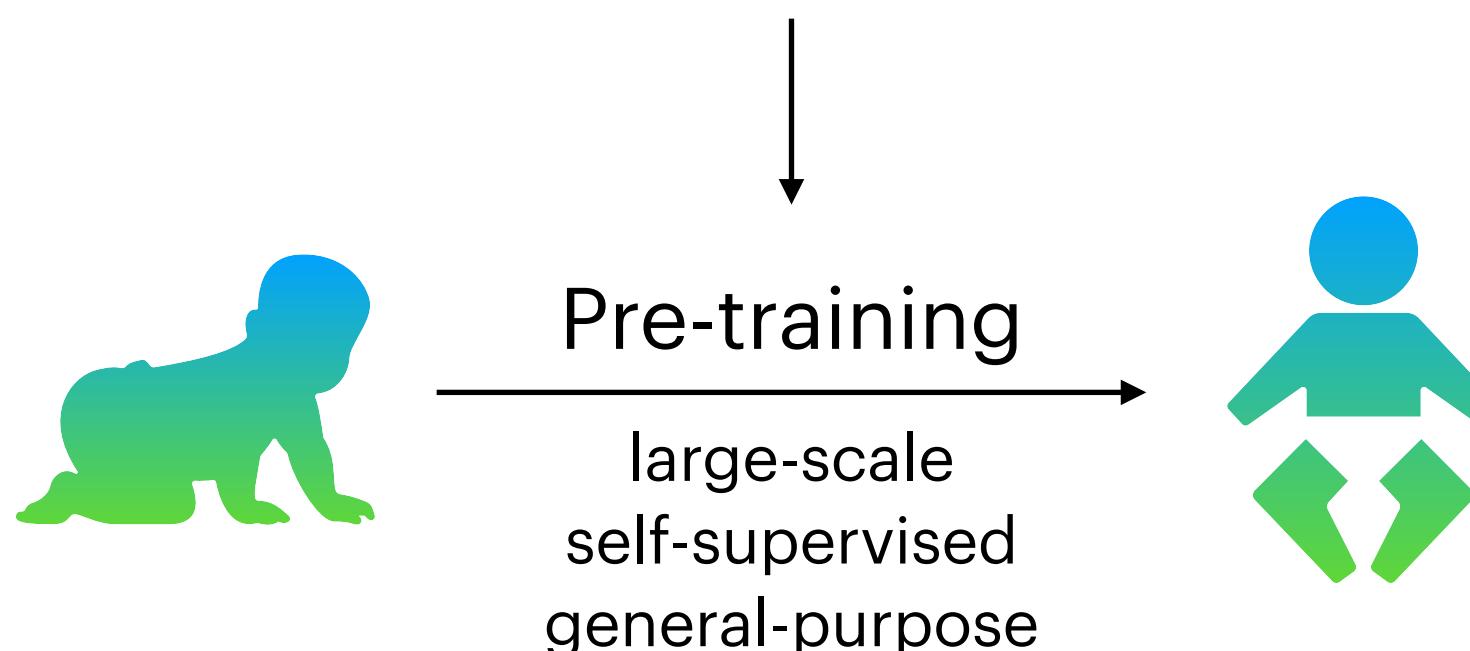
[Source: ImageNet]



[Source: Youtube-8M]



[Source: BRIDGE dataset]



Not done yet...  
The agent needs to keep  
**learning** (new concept, skill, ...)



## Jetson Orin Nano Super Developer Kit

AI Performance	67 INT8 TOPS
GPU	NVIDIA Ampere architecture with 1024 CUDA cores and 32 tensor cores
CPU	6-core Arm® Cortex®-A78AE v8.2 64-bit CPU 1.5MB L2 + 4MB L3
Memory	8GB 128-bit LPDDR5 102 GB/s
Storage	Supports SD card slot and external NVMe
Power	7W–25W

[Source: Nvidia]

Suitable for 3B/4B models  
Inference is fine but training is impractical

Say a 3B model, stored in bf16, training requires

- weights: ~6GB
- gradients: ~6GB
- optimizer states: ~12GB
- activations: ~6GB
- total: ~**30GB** (safer with 32GB)

Far exceeding the capacity of **8GB** Jetson Orin Nano!!

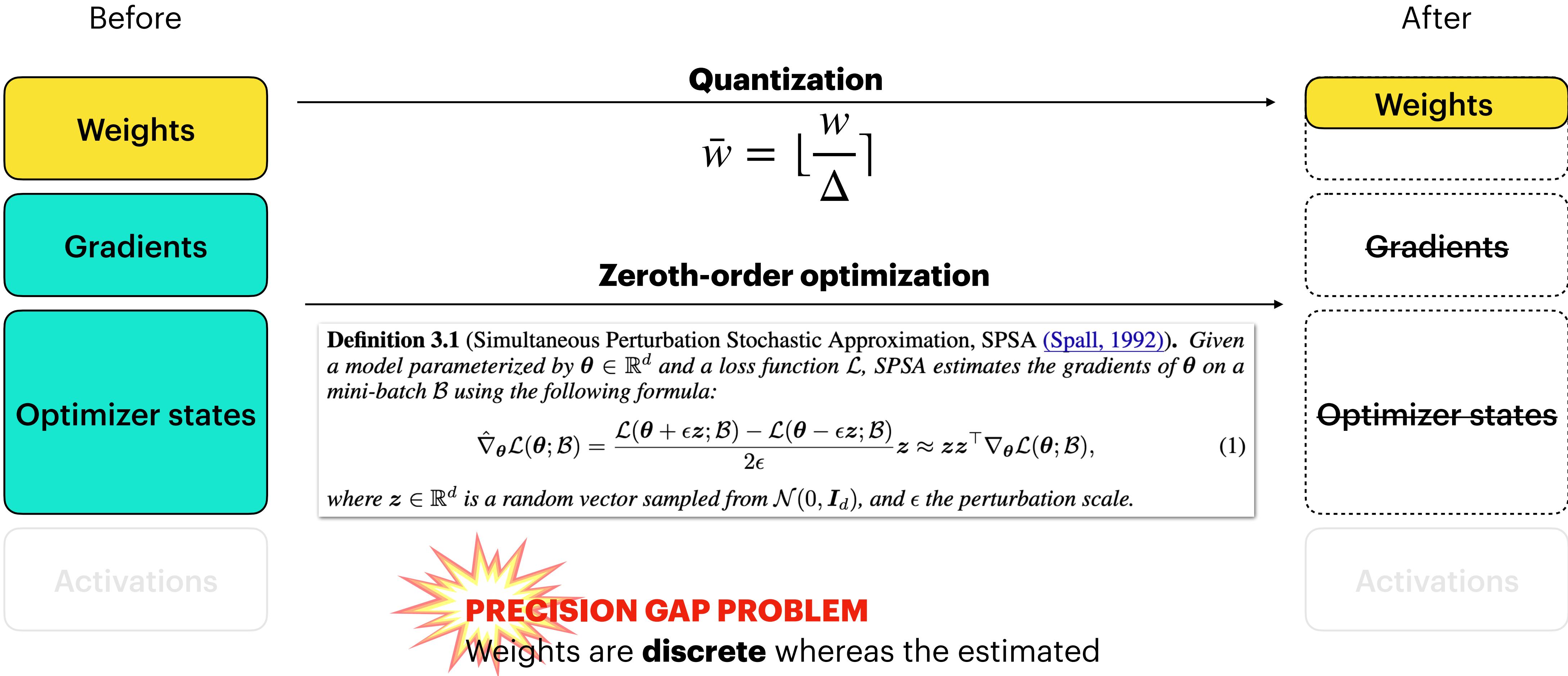
NVIDIA Jetson AGX Orin 64GB? Priced at \$1,999

Nano costs only \$249

Price difference: **\$1,750**

x1000 = **\$1,750,000**, too expensive for mass production

# Trim that memory



# QZO: Quantized Zeroth-order Optimization

A smart design: apply perturbation to the continuous quantization scale!!

**Definition 3.3** (Quantized Simultaneous Perturbation Stochastic Approximation, Q-SPSA). *Given a quantized model with integer parameters  $\bar{\theta} \in \mathbb{R}^d$  and quantization scales  $\Delta$ , and a loss function  $\mathcal{L}$ , Q-SPSA estimates the gradients of  $\Delta$  over a mini-batch  $\mathcal{B}$  using the following formula:*

$$\begin{aligned}\hat{\nabla}_\Delta \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B}) &= \frac{\mathcal{L}((\Delta + \epsilon z) \odot \bar{\theta}; \mathcal{B}) - \mathcal{L}((\Delta - \epsilon z) \odot \bar{\theta}; \mathcal{B})}{2\epsilon} z \\ &\approx z z^\top \nabla_\Delta \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B}),\end{aligned}\tag{5}$$

where  $z \in \mathbb{R}^d$  is a random vector sampled from  $\mathcal{N}(0, I_d)$ ,  $\epsilon$  the perturbation scale, and  $\odot$  the Hadamard product.

This proves that DDC reduces the variance of the gradient estimate and thereby stabilises training

**Theorem 1.** Clipped gradient estimate  $\hat{\nabla}_\Delta \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})$  is an unbiased estimate of the full gradient of loss w.r.t quantization scales  $\nabla_\Delta \mathcal{L}(\Delta \odot \bar{\theta})$ .

Since  $d'^2 \leq d^2$  by definition of DDC in Eq. 6, the following inequality holds:

$$\mathbb{E}[\|\hat{\nabla}_\Delta \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})\|^2] = \mathbb{E}[d'^2 \|z\|^2] \leq \mathbb{E}[d^2 \|z\|^2] = \mathbb{E}[\|\hat{\nabla}_\Delta \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})\|^2]\tag{7}$$

Therefore, the element-wise variance of the clipped gradient estimate has the following derivation:

$$\begin{aligned}Var[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})] &= \mathbb{E}[\|\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})\|^2] - \mathbb{E}[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})]^2 \\ &\leq \mathbb{E}[\|\hat{\nabla}_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})\|^2] - \mathbb{E}[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})]^2 \\ &= Var[\hat{\nabla}_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})] + \mathbb{E}[\hat{\nabla}_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})]^2 - \mathbb{E}[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})]^2 \\ &= Var[\hat{\nabla}_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})] + (\nabla_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}))^2 - \mathbb{E}[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})]^2\end{aligned}\tag{8}$$

By Theorem 1,  $Var[\hat{\nabla}_{\Delta_k} \mathcal{L}'(\Delta \odot \bar{\theta}; \mathcal{B})] \leq Var[\hat{\nabla}_{\Delta_k} \mathcal{L}(\Delta \odot \bar{\theta}; \mathcal{B})]$  holds almost surely.

directional derivative clipping (DDC)

$$d' = \begin{cases} C, & \text{if } d > C \\ d, & \text{if } d \in [-C, C] \\ -C, & \text{if } d < -C \end{cases}$$

Table 1: Experiments based on OPT-6.7B, Llama-2-7B, and Llama-3.1-8B. Zero-Shot and Zero-Shot-Q serve as the lower-bound, while Fine-tuning (with SGD) is the upper-bound. QZO works well across different model architectures on all datasets, with a memory footprint significantly lower than MeZO and Fine-tuning.

		Model Precision	Memory Profiling	SST-2	Classification			Generation SQuAD
					RTE	CB	BoolQ	
OPT-6.7B	Fine-tuning	16 bits	26.8GB	95.4	79.8	73.2	69.6	77.6
	Zero-Shot	16 bits	-	61.2	55.2	51.8	59.5	36.5
	Zero-Shot-Q	4 bits	-	60.1	53.8	51.8	59.1	35.9
	MeZO	16 bits	14.8GB	93.0	64.6	67.9	66.8	79.6
	QZO	4 bits	4.8GB	87.6	61.7	67.9	66.4	78.5
Llama-2-7B	Fine-tuning	16 bits	26.0GB	92.8	63.2	60.7	75.0	83.7
	Zero-Shot	16 bits	-	58.1	61.7	32.1	66.0	55.6
	Zero-Shot-Q	4 bits	-	58.5	53.4	35.7	64.6	53.6
	MeZO	16 bits	14.8GB	83.5	58.1	67.9	69.6	80.7
	QZO	4 bits	5.0GB	90.0	59.2	69.6	68.2	85.5
Llama-3-8B	Fine-tuning	16 bits	31.9GB	93.7	71.5	62.5	83.4	84.9
	Zero-Shot	16 bits	-	59.6	45.8	46.4	66.1	64.8
	Zero-Shot-Q	4 bits	-	58.7	50.2	37.5	65.0	59.2
	MeZO	16 bits	20.5GB	92.5	70.0	91.1	83.4	86.9
	QZO	4 bits	6.3GB	93.0	66.8	69.6	78.2	88.3

- sentiment analysis
- logical inference
- reading comprehension
- QA

# Fine-tuning **13B LLM** within a single 24GB GPU



Nvidia RTX 4090 24GB

	Model Precision	Memory Profiling	Classification				Generation SQuAD
			SST-2	RTE	CB	BoolQ	
Llama-2-13B	Zero-Shot-Q QZO	2 bits 2 bits	- 5.78GB	57.6 80.5	53.1 54.5	46.4 55.4	69.2 70.2
							55.4 59.4

Under such an extreme quantization setting, QZO's gains are very encouraging

It's worth noting that a 13B LLM would normally require **65GB** of memory for fine-tuning—QZO did it just **under 6GB!**

# Fine-tuning SD 3.5 Large within a single 24GB GPU



Nvidia RTX 4090 24GB

QZO's images look slightly better than zero-shot. Not perfect but encouraging

Stable Diffusion 3.5 Large requires at least **87GB** of memory for fine-tuning

QZO uses only **12.4GB**



# Summary

- Multimodal foundation models
  - GPT, next-token prediction
  - instruction tuning, scaling, data quality, alignment
- Pathway to multimodal agents
  - reasoning: think fast and slow
  - grounding: grounded chain-of-thought
  - Awareness: streaming video, data-driven learning
  - Adaptation: on-device learning, zeroth-order optimization