(Visual) Instruction Tuning

Visual Instruction Tuning

One of the core aspirations in artificial intelligence is to develop a ***general-purpose assistant*** that can effectively follow multi-modal vision-and-language instructions

Visual Instruction Tuning

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User: What is unusual about this image?

BLIP-2: a man is sitting on the back of a yellow cab

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.

Language Modeling ≠ Following Human Instructions

GPT-3

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.
```

Language models are not aligned with user intents [Ouyang et al., 2022].

[Training language models to follow instructions with human feedback, Ouyang et al. 2022]

Language Modeling ≠ Following Human Instructions

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

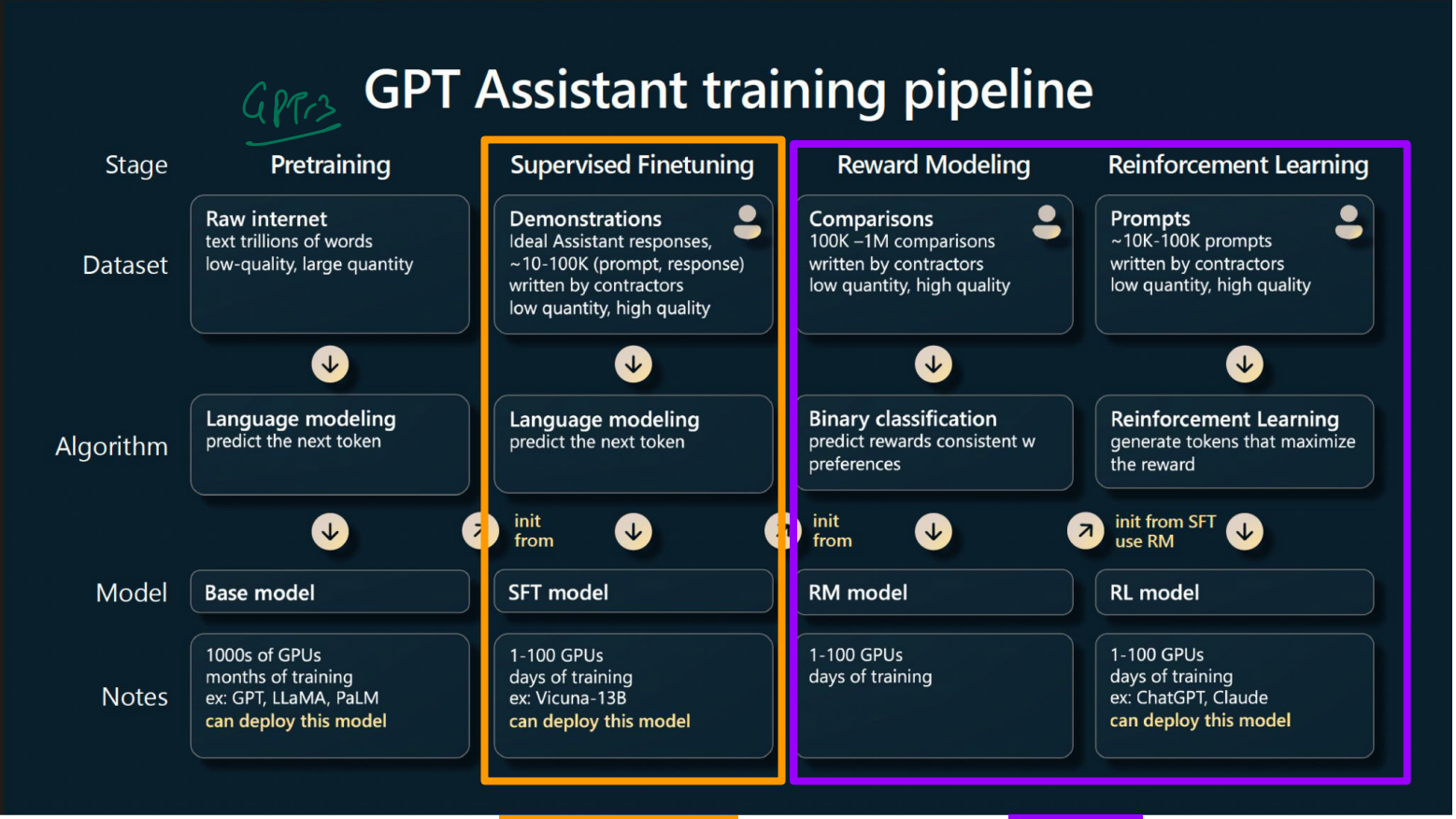
COMPLETION Human

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

Language models are not aligned with user intents [Ouyang et al., 2022].

[Training language models to follow instructions with human feedback, Ouyang et al. 2022]

How do we make LMs aligned with our intents that are articulated in language?



Instruction Fine-Tuning

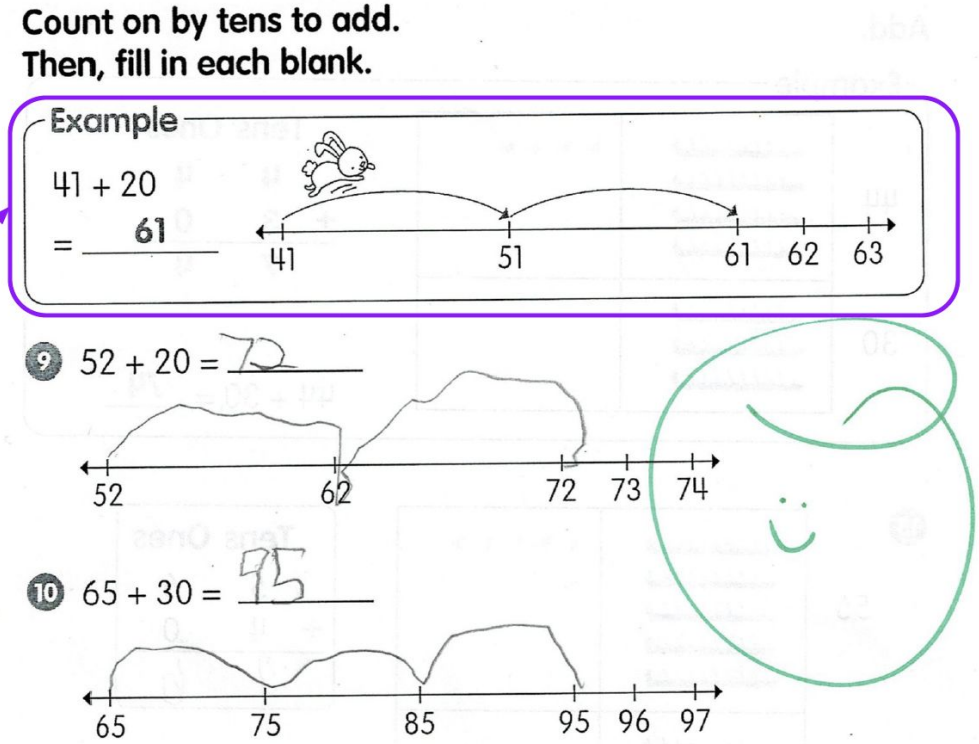
RLHF

Source: State of GPT, Karpathy

What is Instruction Fine-Tuning?

How do humans learn?

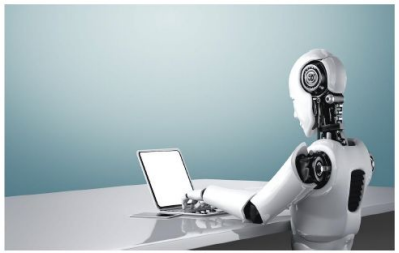
- Task instruction
  - how to “count”, and
  - where to “fill”
- Very few examples
  - only one here



Typical human learning

- Detailed instruction
- Very few examples (≤ 10)

Mainstream machine learning



- No instructions
- Large training sets
- Even few-shot learning often uses 100s of examples

Basic Premise

NLP tasks can be described via natural language instructions, such as

"*Is the sentiment of this movie review positive or negative*?"

or

"*Translate 'how are you' into Chinese.*"

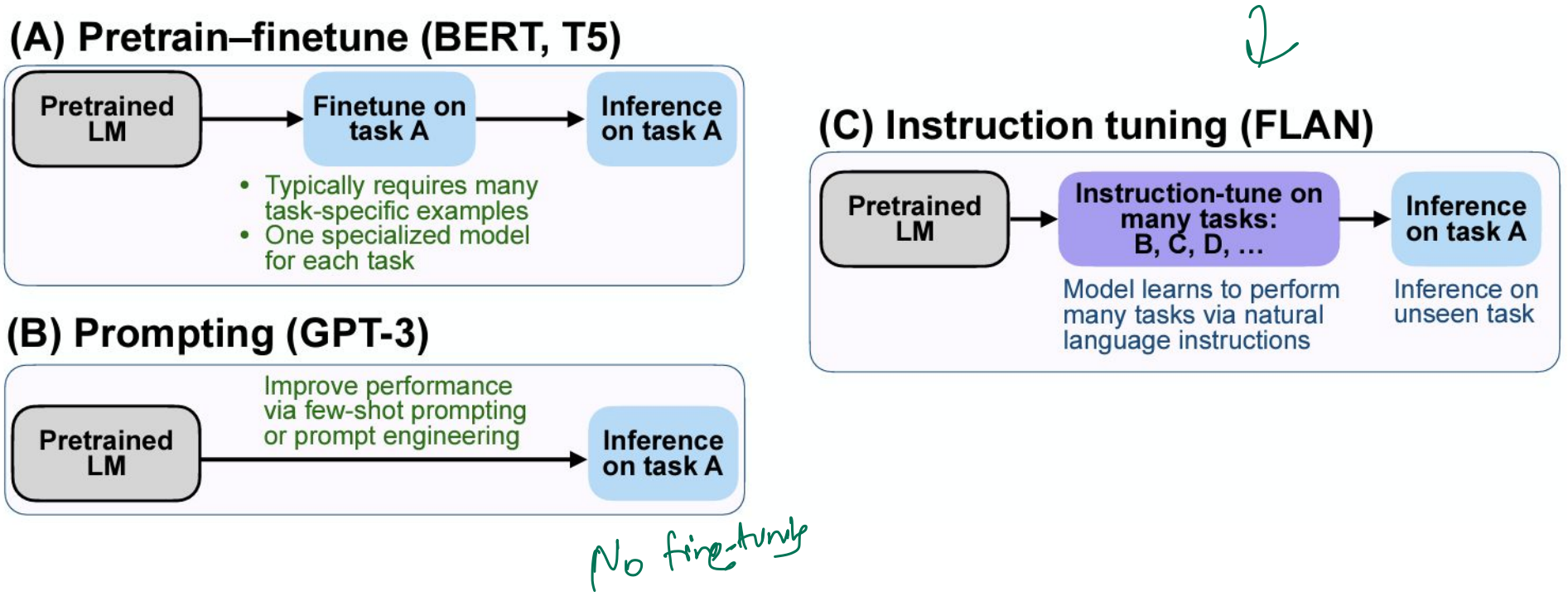
Why define tasks in natural language?

While the current dominant paradigm (supervised learning with task-specific labeled examples) has been successful in building task-specific models, ***such models can't generalize to unseen tasks***; for example, a model that is supervised to solve questions cannot solve a classification task.

We hypothesize that ***a model equipped with understanding and reasoning with natural language instructions should be able to generalize*** to any task that can be defined in terms of natural language.

Source: https://github.com/allenai/natural-instructions

How is instruction tuning a different paradigm?



Finetuned Language Models Are Zero-Shot Learners, ICLR'22.

An example of instruction

Sample Extended Instruction

- **Definition:** This task involves creating answers to complex questions, from a given passage. Answering these questions, typically involve understanding multiple sentences. Make sure that your answer has the same type as the "answer type" mentioned in input. The provided "answer type" can be of any of the following types: "span", "date", "number". A "span" answer is a continuous phrase taken directly from the passage or question. You can directly copy-paste the text from the passage or the question for span type answers. If you find multiple spans, please add them all as a comma separated list. Please restrict each span to five words. A "number" type answer can include a digit specifying an actual value. For "date" type answers, use DD MM YYYY format e.g. 11 Jan 1992. If full date is not available in the passage you can write partial date such as 1992 or Jan 1992.
- **Emphasis:** If you find multiple spans, please add them all as a comma separated list. Please restrict each span to five words.
- **Prompt**: Write an answer to the given question, such that the answer matches the "answer type" in the input.
  Passage: {passage}
  Question: { question }

Figure 12.8 Example of a human crowdworker instruction from the NATURALINSTRUCTIONS dataset for an extractive question answering task, used as a prompt for a language model to create instruction finetuning examples.

https://web.stanford.edu/~jurafsky/slp3/

Creating instructions at scale

Few-Shot Learning for QA

Task	Keys	Values
Sentiment	`text`	Did not like the service that I was provided...
	`label`	0
	`text`	It sounds like a great plot, the actors are first grade, and...
	`label`	1
NLI	`premise`	No weapons of mass destruction found in Iraq yet.
	`hypothesis`	Weapons of mass destruction found in Iraq.
	`label`	2
	`premise`	Jimmy Smith... played college football at University of Colorado.
	`hypothesis`	The University of Colorado has a college football team.
	`label`	0
Extractive Q/A	`context`	Beyoncé Giselle Knowles-Carter is an American singer...
	`question`	When did Beyonce start becoming popular?
	`answers`	{ `text`: ['in the late 1990s'], `answer_start`: 269 }

Figure 12.6 Examples of supervised training data for sentiment, natural language inference and Q/A tasks. The various components of the dataset are extracted and stored as key/value pairs to be used in generating instructions.

https://web.stanford.edu/~jurafsky/slp3/

Creating instructions at scale

Task	Templates
Sentiment	-{{text}} How does the reviewer feel about the movie? -The following movie review expresses what sentiment? {{text}} -{{text}} Did the reviewer enjoy the movie?
Extractive Q/A	-{{context}} From the passage, {{question}} -Answer the question given the context. Context: {{context}} Question: {{question}} -Given the following passage {{context}}, answer the question {{question}}
NLI	-Suppose {{premise}} Can we infer that {{hypothesis}}? Yes, no, or maybe? -{{premise}} Based on the previous passage, is it true that {{hypothesis}}? Yes, no, or maybe? -Given {{premise}} Should we assume that {{hypothesis}} is true? Yes,no, or maybe?

Figure 12.7 Instruction templates for sentiment, Q/A and NLI tasks.

Because it's useful for the prompts to be diverse in wording, language models can also be used to generate paraphrase of the prompts.

https://web.stanford.edu/~jurafsky/slp3/

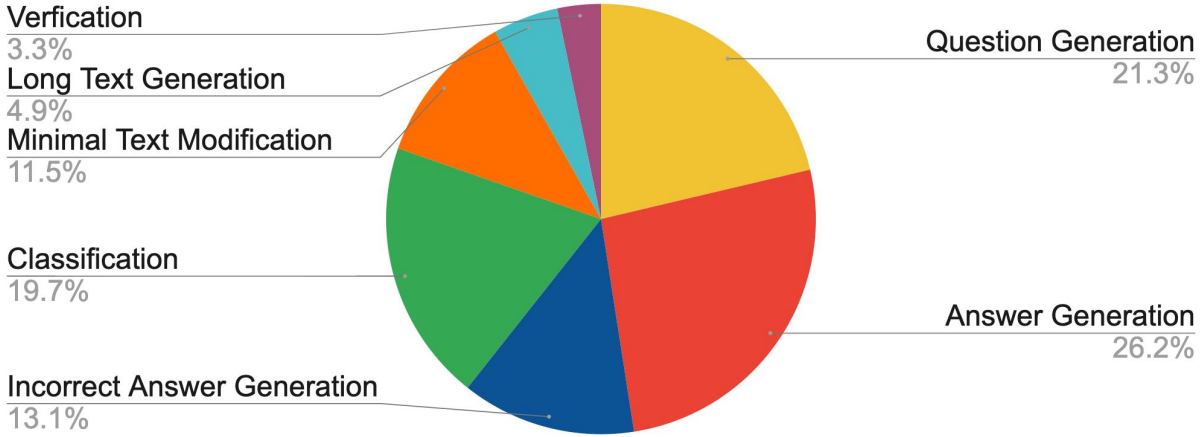
Natural Instructions: Statistics

- **Tasks**: ~61
- **Instances**: ~193k
- **Categories**: ~7
- **Diverse Reasoning Skills**:

  E.g. Numerical Reasoning, Coreference Resolution, Commonsense Reasoning, Multihop Reasoning.

- **Diverse Domains:**

  E.g. Sports, History, News, Conversations, Geography, NFL games, Captions, Maths.



Cross-Task Generalization via Natural Language Crowdsourcing Instructions. **ACL 22**

Finetune on many tasks ("instruction-tuning")

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.

How would you accomplish this goal?

OPTIONS:

-Keep stack of pillow cases in fridge.

-Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:

The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...

Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

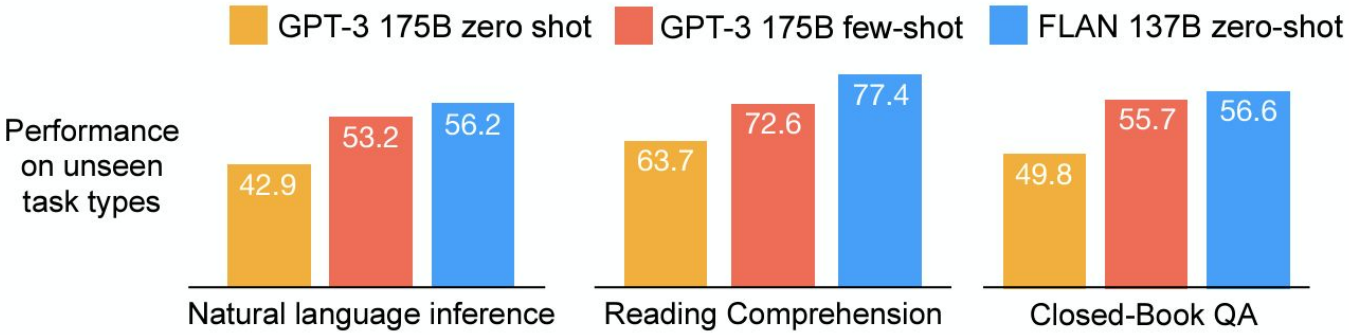
Does the premise entail the hypothesis?

OPTIONS:

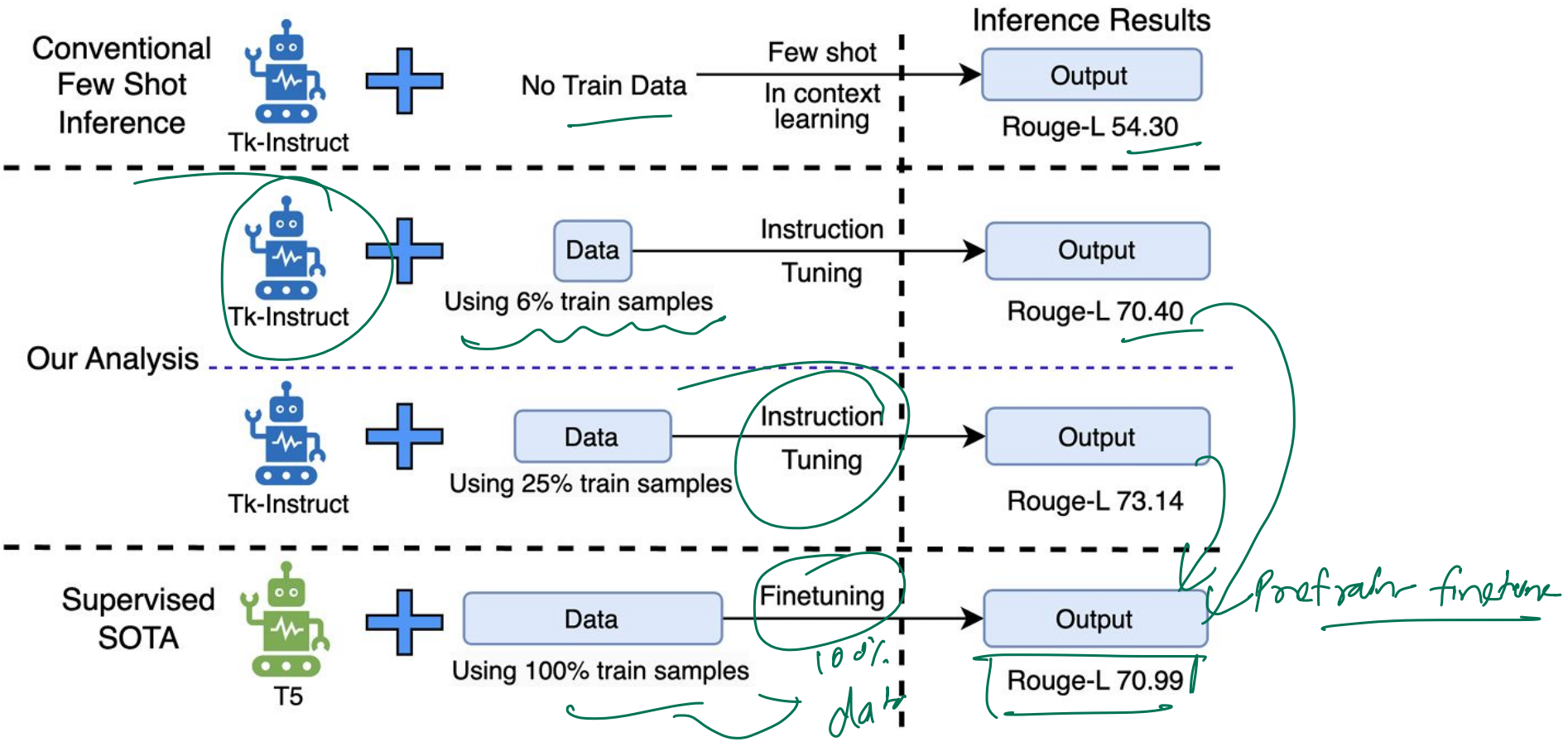
-yes -it is not possible to tell -no

FLAN Response

It is not possible to tell



Finetuned Language Models Are Zero-Shot Learners, ICLR'22.



Instruction Tuned Models are quick Learners

Himanshu Gupta*, Saurabh Arjun Sawant*, Swaroop Mishra, Santosh Mashetty, Mutsumi Nakamura, Arindam Mitra, Chitta Baral

Flan Models

- Instruction fine-tuning can be done on various models (PaLM, T5, etc.)
  - Flan-T5 models publicly available

Params	Model	Arhitecture	pre-training Objective	Pretrain FLOPs	Finetune FLOPs	% Finetune Compute
80M	Flan-T5-Small	encoder-decoder	span corruption	1.8E+20	2.9E+18	1.6%
250M	Flan-T5-Base	encoder-decoder	span corruption	6.6E+20	9.1E+18	1.4%
780M	Flan-T5-Large	encoder-decoder	span corruption	2.3E+21	2.4E+19	1.1%
3B	Flan-T5-XL	encoder-decoder	span corruption	9.0E+21	5.6E+19	0.6%
11B	Flan-T5-XXL	encoder-decoder	span corruption	3.3E+22	7.6E+19	0.2%
8B	Flan-PaLM	decoder-only	causal LM	3.7E+22	1.6E+20	0.4%
62B	Flan-PaLM	decoder-only	causal LM	2.9E+23	1.2E+21	0.4%
540B	Flan-PaLM	decoder-only	causal LM	2.5E+24	5.6E+21	0.2%
62B	Flan-cont-PaLM	decoder-only	causal LM	4.8E+23	1.8E+21	0.4%
540B	Flan-U-PaLM	decoder-only	prefix LM + span corruption	2.5E+23	5.6E+21	0.2%

Table 2: Across several models, instruction finetuning only costs a small amount of compute relative to pre-training. T5: Raffel et al. (2020). PaLM and cont-PaLM (also known as PaLM 62B at 1.3T tokens): Chowdhery et al. (2022). U-PaLM: Tay et al. (2022b).

Chung et al. (2022)

Visual Instruction Tuning

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User: What is unusual about this image?

BLIP-2: a man is sitting on the back of a yellow cab

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.

How to create the instruction tuning dataset?

From the paper: We collect 158K unique language-image instruction-following samples in total, including 58K in conversations, 23K in detailed description, and 77k in complex reasoning, respectively.

But where do you collect such data?

2023

ChatGPT

GPT-4

https://arxiv.org/pdf/2304.08485

How to create the instruction tuning dataset?

From the paper: We collect 158K unique language-image instruction-following samples in total, including 58K in conversations, 23K in detailed description, and 77k in complex reasoning, respectively.

But where do you collect such data?

Leverage ChatGPT/GPT-4 for multimodal instruction-following data collection, based on the ***widely existing image-text paired data***.

Same as BLIP-2

But how exactly? These models were not taking image input

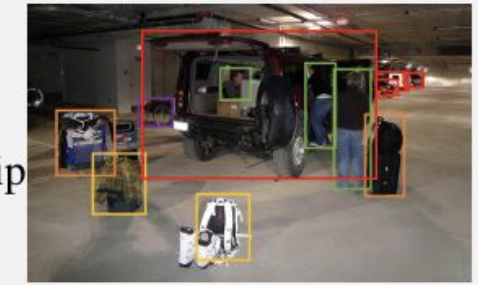
https://arxiv.org/pdf/2304.08485

Gathering synthetic data for instruction tuning

In order to encode an image into its visual features to prompt a text-only GPT, we use two types of symbolic representations:
(i) Captions typically describe the visual scene from various perspectives;
(ii) Bounding boxes usually localize the objects in the scene, and each box encodes the object concept and its spatial location.

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.
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

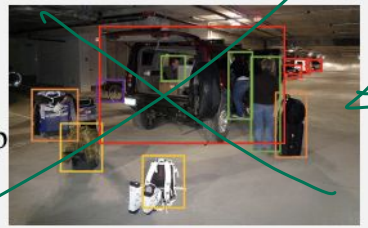
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Gathering synthetic data for instruction tuning

Generate three types of instruction following data by (few-shot prompting)* 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.
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], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Gathering synthetic data for instruction tuning

Generate three types of instruction following data by prompting GPT-4

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.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

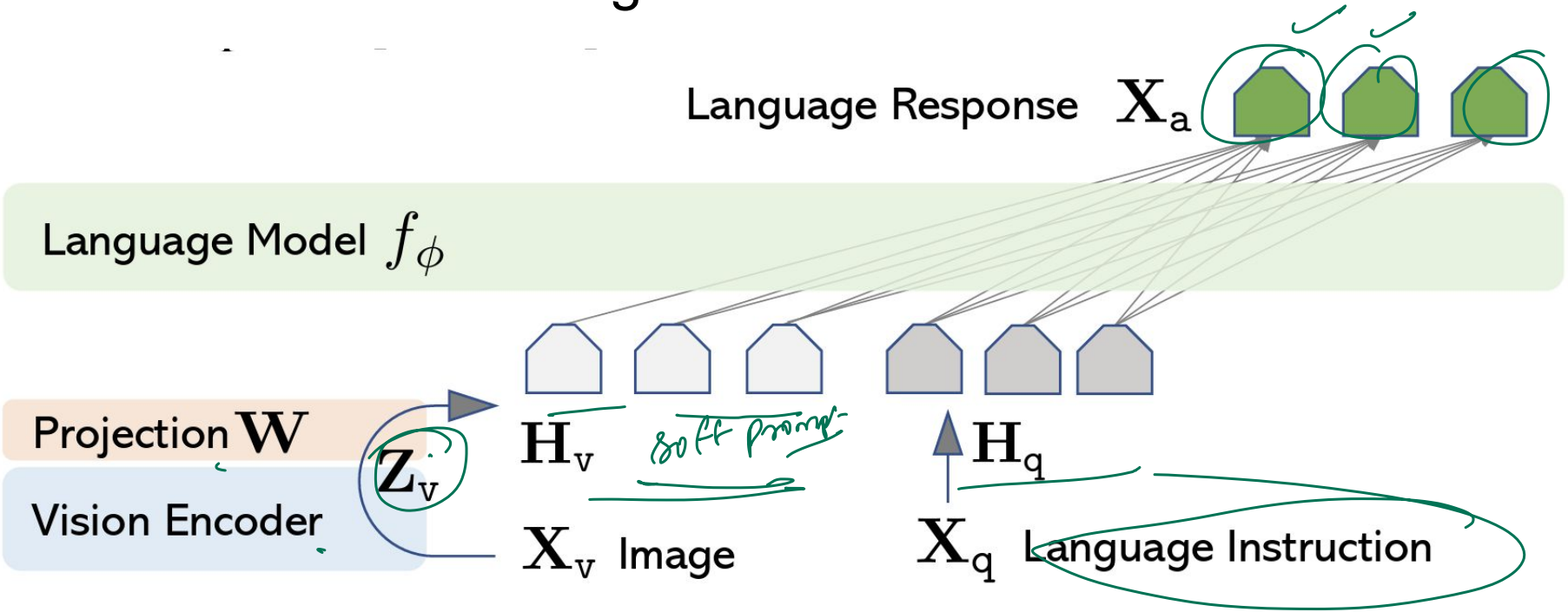
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

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 to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Visual Instruction Tuning



For an input image X_v, we consider the pre-trained CLIP visual encoder ViT-L/14, which provides the visual feature $Z_v = g(X_v)$

We apply a trainable projection matrix W to convert Z_v into language embedding tokens H_v, which have the same dimensionality as the word embedding space in the language model

$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v,$$

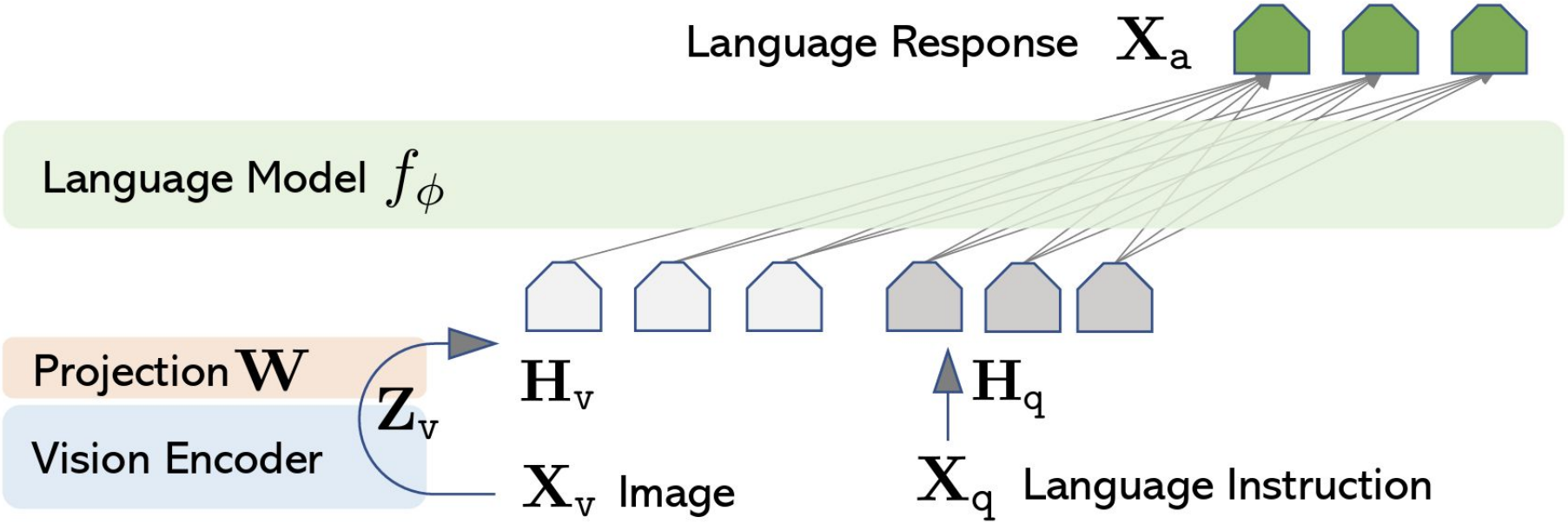
Data Format

For each image $\mathbf{X}_\mathrm{v}$, we generate multi-turn conversation data $(\mathbf{X}_\mathrm{q}^1, \mathbf{X}_\mathrm{a}^1, \cdots, \mathbf{X}_\mathrm{q}^T, \mathbf{X}_\mathrm{a}^T)$, where T is the total number of turns. We organize them as a sequence, by treating all answers as the assistant's response, and the instruction $\mathbf{X}_\mathrm{instruct}^t$ at the t-th turn as:

$$\mathbf{X}_\mathrm{instruct}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_\mathrm{q}^1, \mathbf{X}_\mathrm{v}] \text{ or } [\mathbf{X}_\mathrm{v}, \mathbf{X}_\mathrm{q}^1], & \text{the first turn } t = 1 \\ \mathbf{X}_\mathrm{q}^t, & \text{the remaining turns } t > 1 \end{cases}$$

We perform instruction-tuning of the LLM on the prediction tokens, using its original auto-regressive training objective.

Visual Instruction Tuning: Training



$$p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_v, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{a,<i})$$