

CSCE 689: Special Topics in Trustworthy NLP

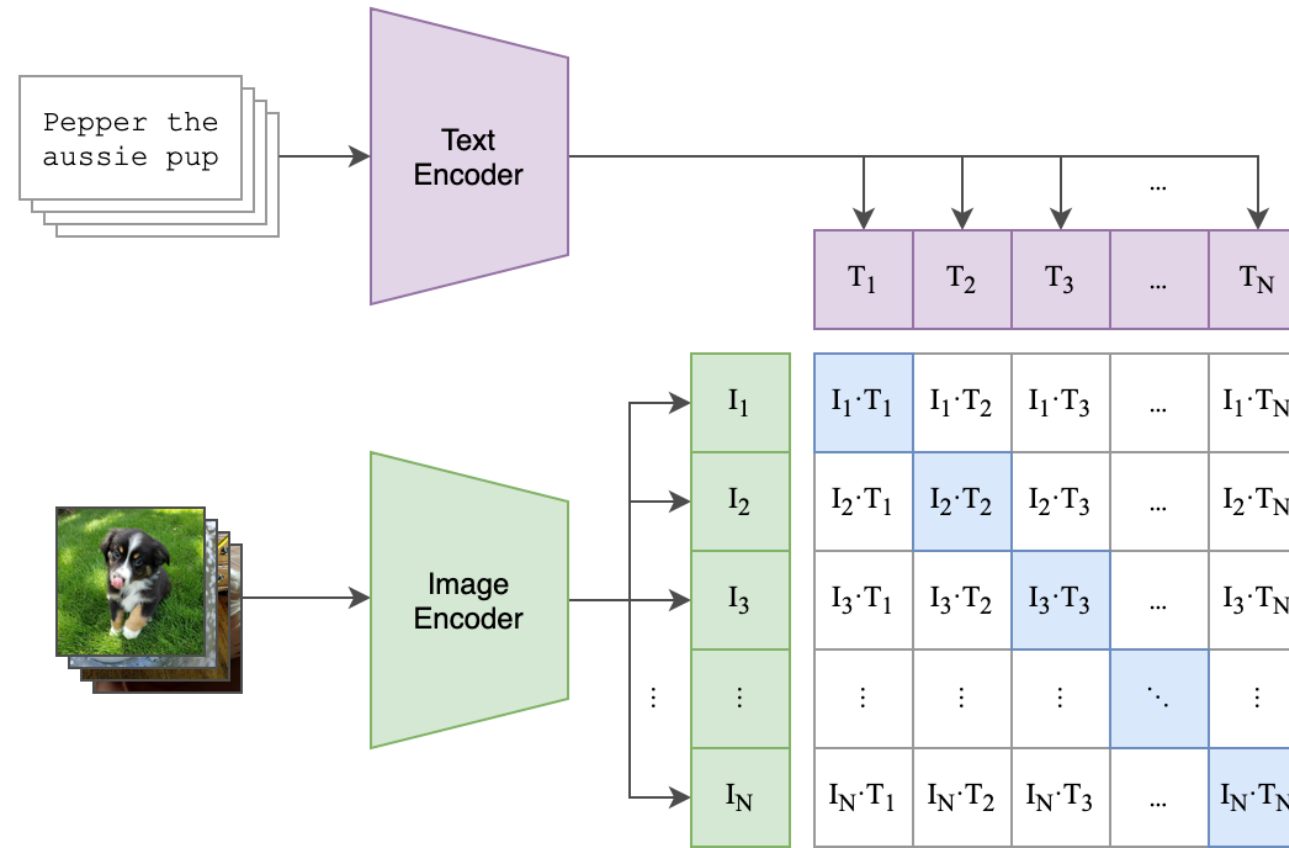
Lecture 8: Human Preference Alignment

Kuan-Hao Huang
khhuang@tamu.edu



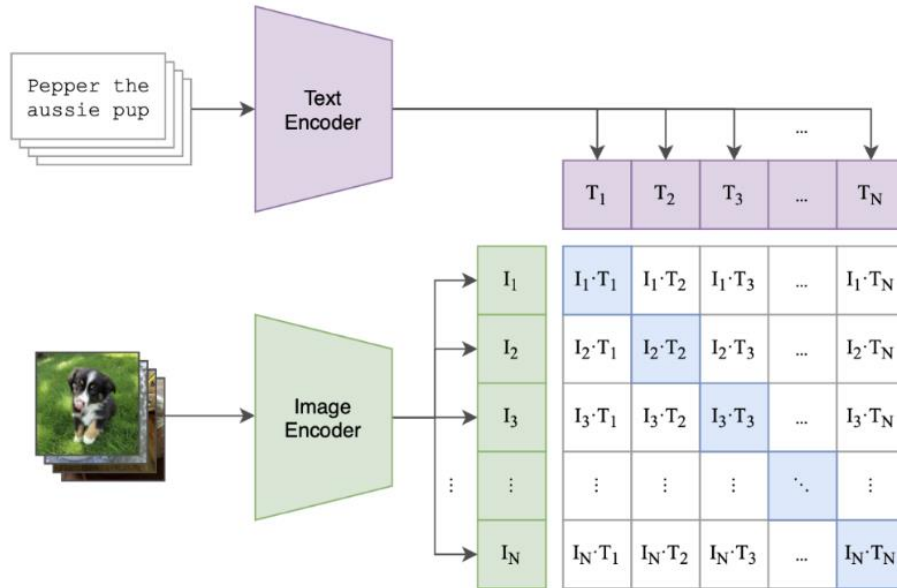
(Some slides adapted from Graham Neubig, Jesse Mu, and Hung-Yi Lee)

Recap: CLIP: Contrastive Language-Image Pre-Training

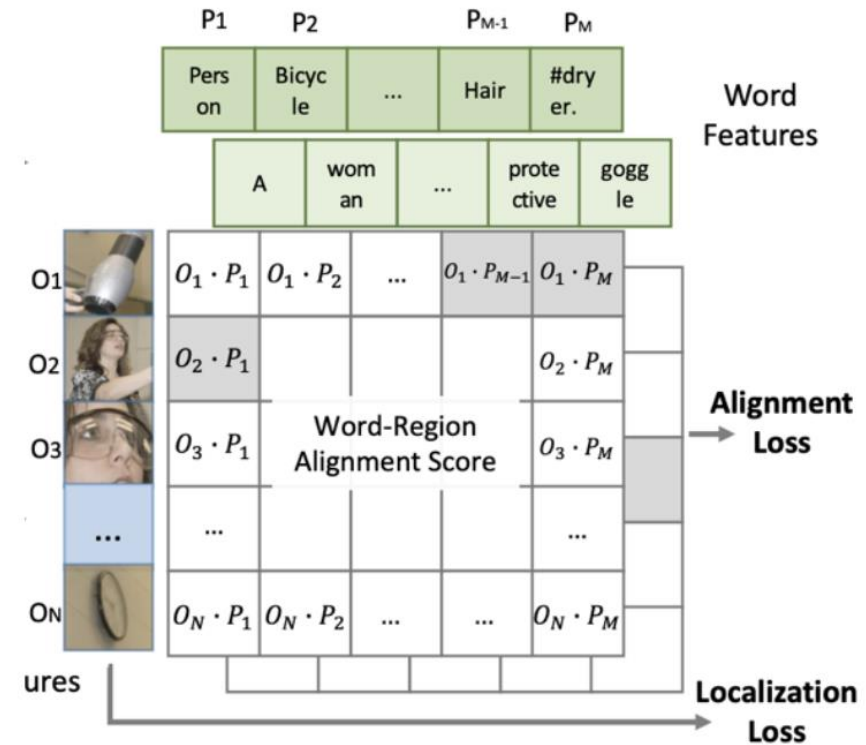


GLIP: Grounded Language-Image Pre-training

CLIP: capture information for whole image



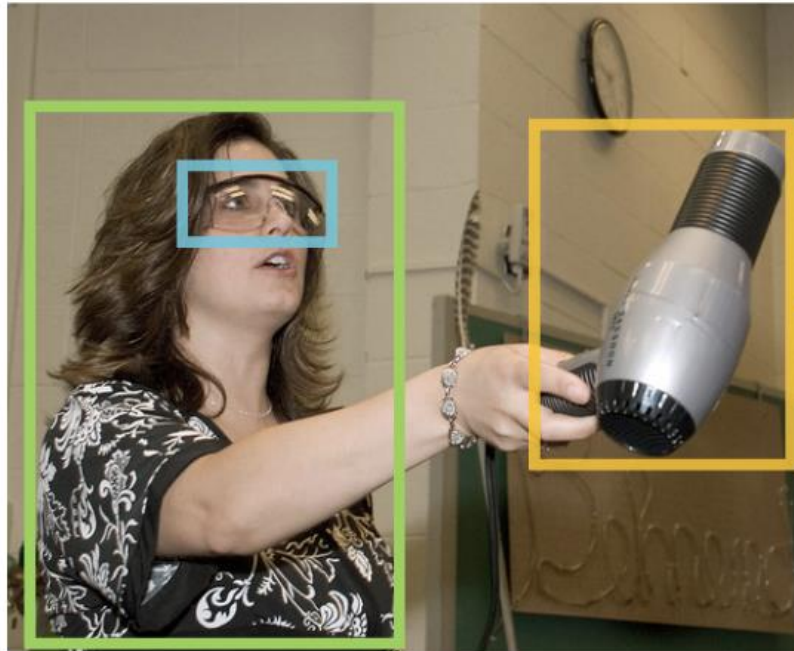
GLIP: capture information more for objects/entities



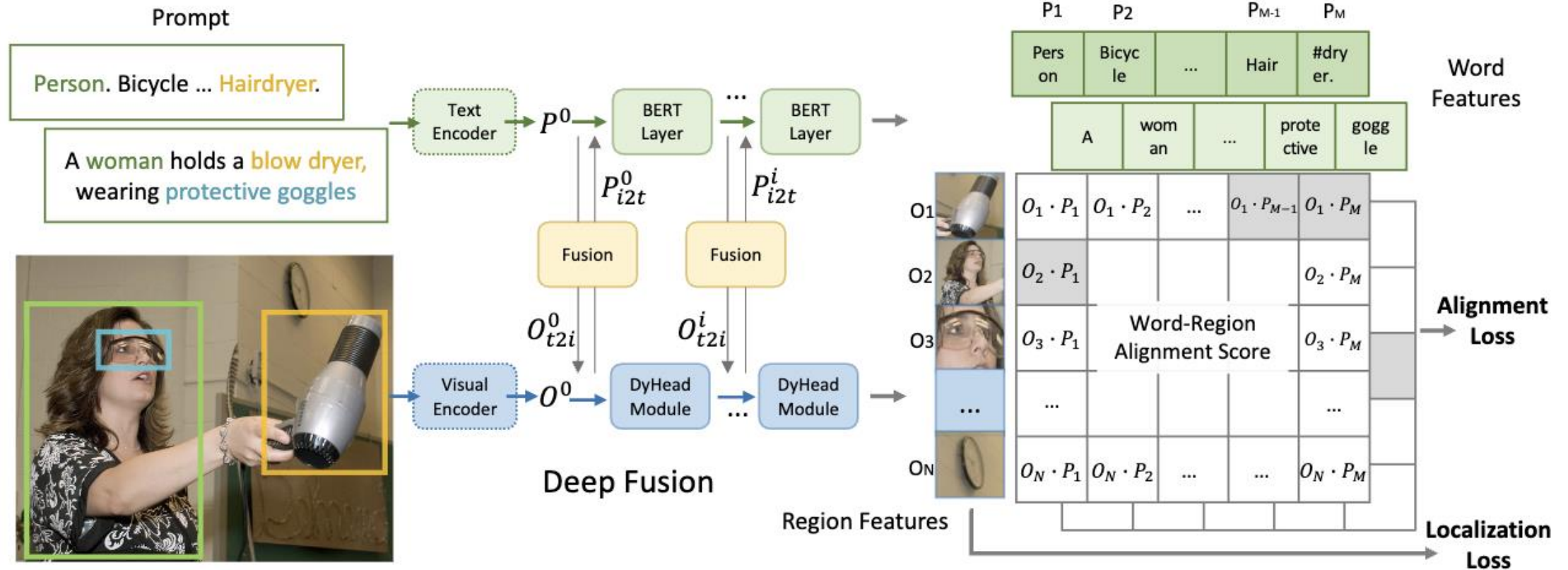
Object Detection and Text Grounding

Person. Bicycle ... Hairdryer.

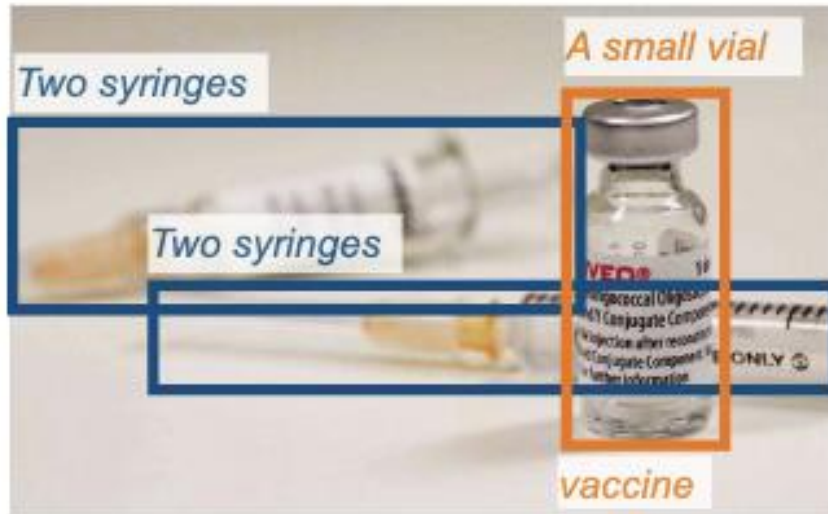
A woman holds a blow dryer,
wearing protective goggles



GLIP: Grounded Language-Image Pre-training



Grounding Results



Two syringes and a small vial of vaccine.



playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

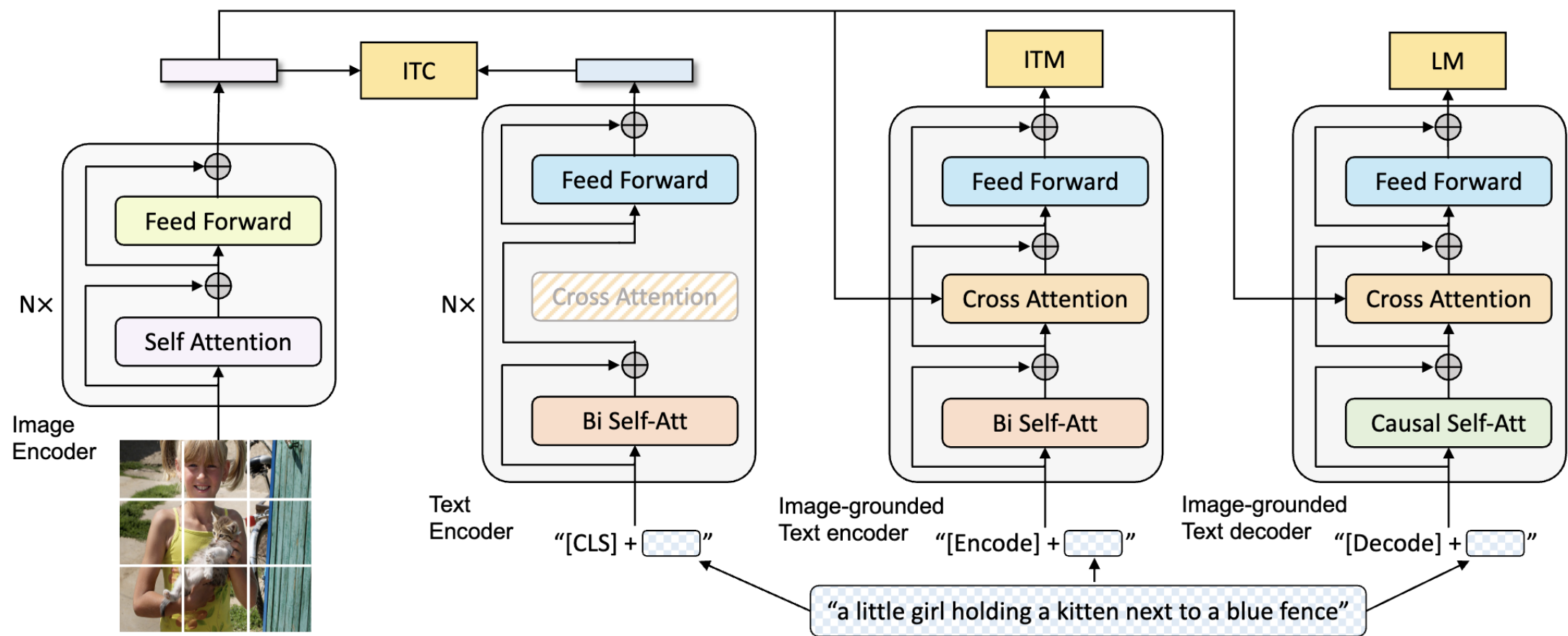
Zero-Shot Grounding

Model	Backbone	MiniVal [23]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [23]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [23]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [15]	RN50	-	-	-	-	12.3	24.3	32.4	25.4
GLIP-T (A)	Swin-T	14.2	13.9	23.4	18.5	6.0	8.0	19.4	12.3
GLIP-T (B)	Swin-T	13.5	12.8	22.2	17.8	4.2	7.6	18.6	11.3
GLIP-T (C)	Swin-T	17.7	19.5	31.0	24.9	7.5	11.6	26.1	16.5
GLIP-T	Swin-T	20.8	21.4	31.0	26.0	10.1	12.5	25.5	17.2
GLIP-L	Swin-L	28.2	34.3	41.5	37.3	17.1	23.3	35.4	26.9

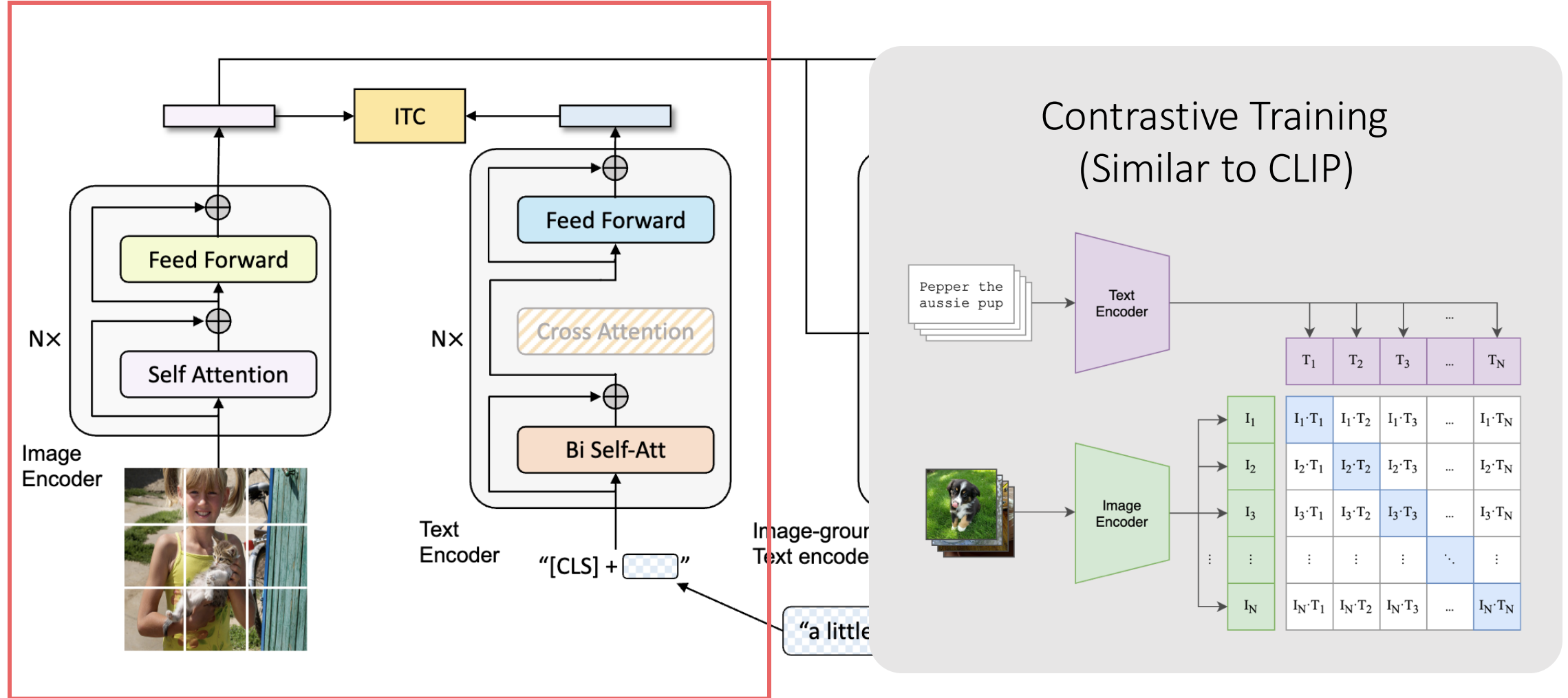
Encoder-Only vs. Encoder-Decoder

- Encoder-only
 - CLIP, GLIP, DesCo, etc.
 - Better for image-text retrieval
- Encoder-decoder
 - Better for generation

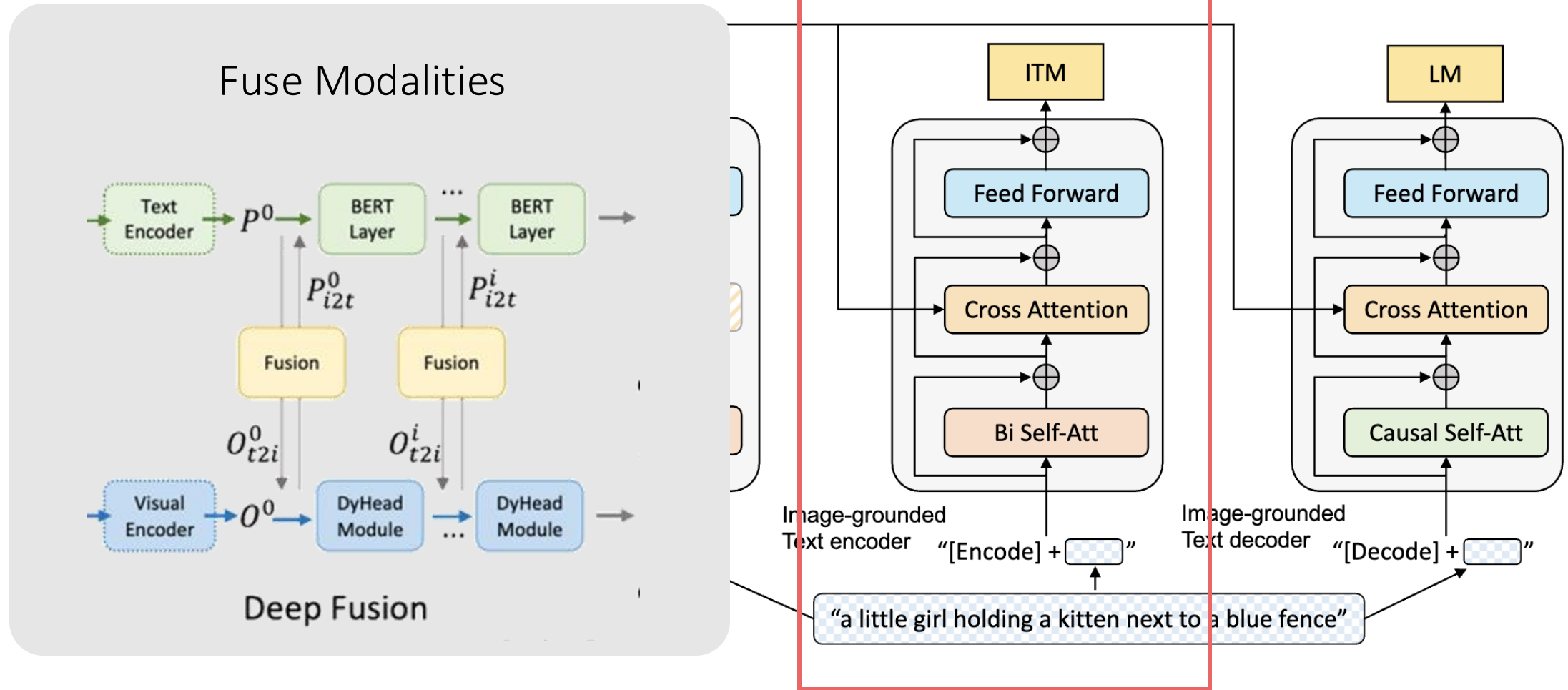
BLIP: Bootstrapping Language-Image Pre-training



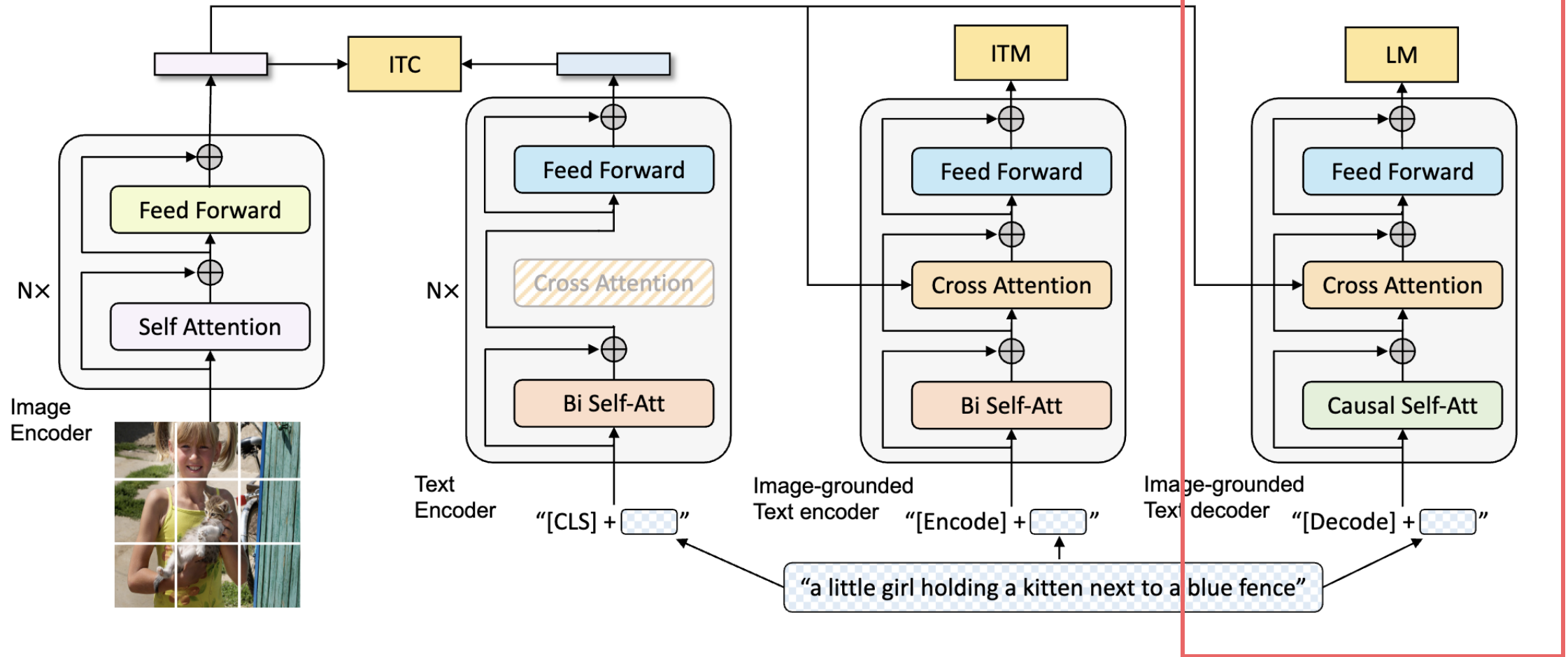
A Unified Framework



A Unified Framework



A Unified Framework



Zero-Shot Image-Text Retrieval

Method	Pre-train # Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	100.0	84.9	96.7	98.3
BLIP	129M	96.0	99.9	100.0	85.0	96.8	98.6
BLIP _{CapFilt-L}	129M	96.0	99.9	100.0	85.5	96.8	98.7
BLIP _{ViT-L}	129M	96.7	100.0	100.0	86.7	97.3	98.7

Image Captioning

Method	Pre-train #Images	NoCaps validation								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall		B@4	C
		C	S	C	S	C	S	C	S		
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL [†] (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
LEMON _{base} [†] (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON _{base} [†] (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	40.3	133.3
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
BLIP _{CapFilt-L}	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7	39.7	133.3
LEMON _{large} [†] (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
SimVLM _{huge} (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP _{ViT-L}	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Visual Question Answering

Visual Question Answering

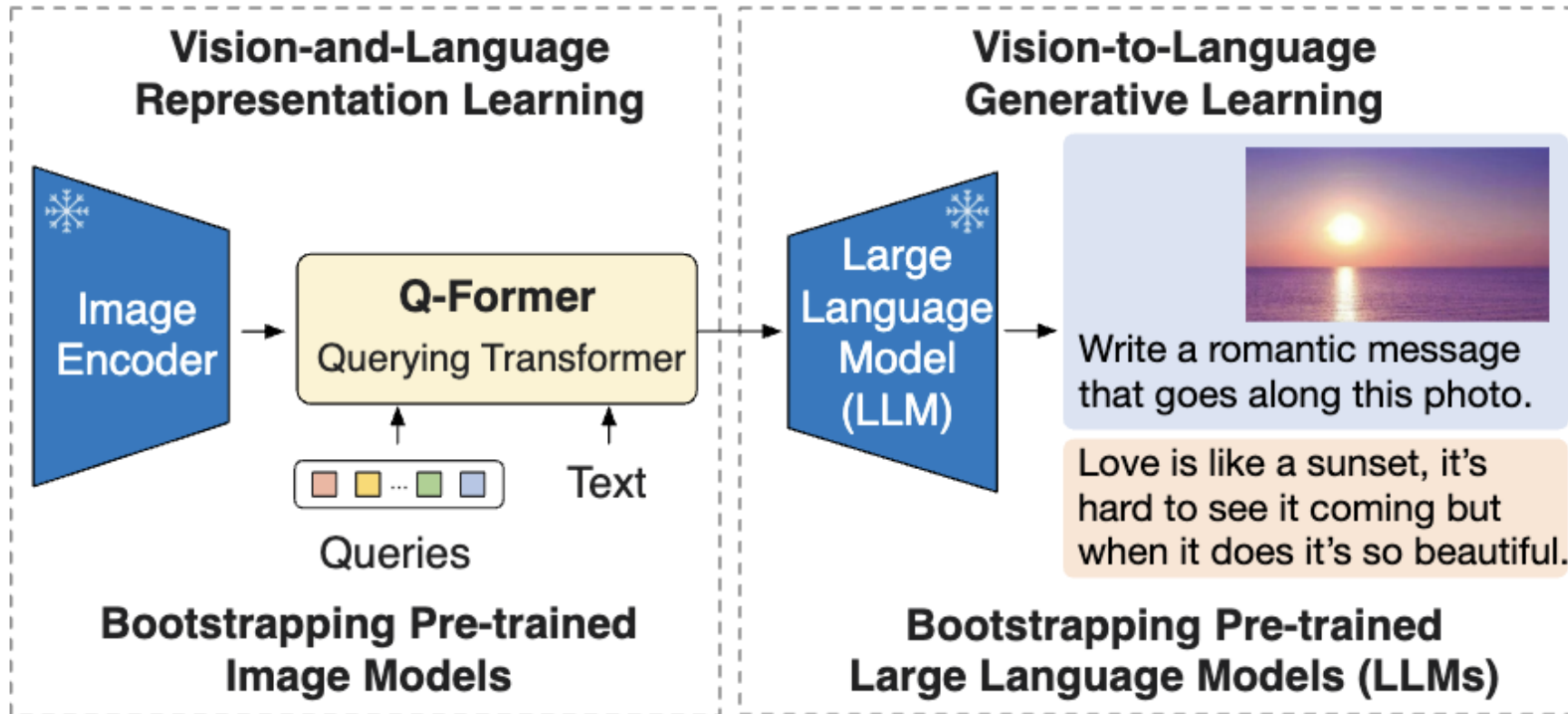


Does it appear to be rainy?


Does this person have 20/20 vision?

Method	Pre-train #Images	VQA		NLVR ²	
		test-dev	test-std	dev	test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM _{base} [†]	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	82.67	82.30
BLIP	129M	78.24	78.17	82.48	83.08
BLIP _{CapFilt-L}	129M	78.25	78.32	82.15	82.24

BLIP-2: Frozen Image Encoders and Large Language Models




Instructed Zero-Shot Image-to-Text Generation




Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.




Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shi Huang to protect the emperor's capital from invaders from the north.



Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.



Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.




What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.



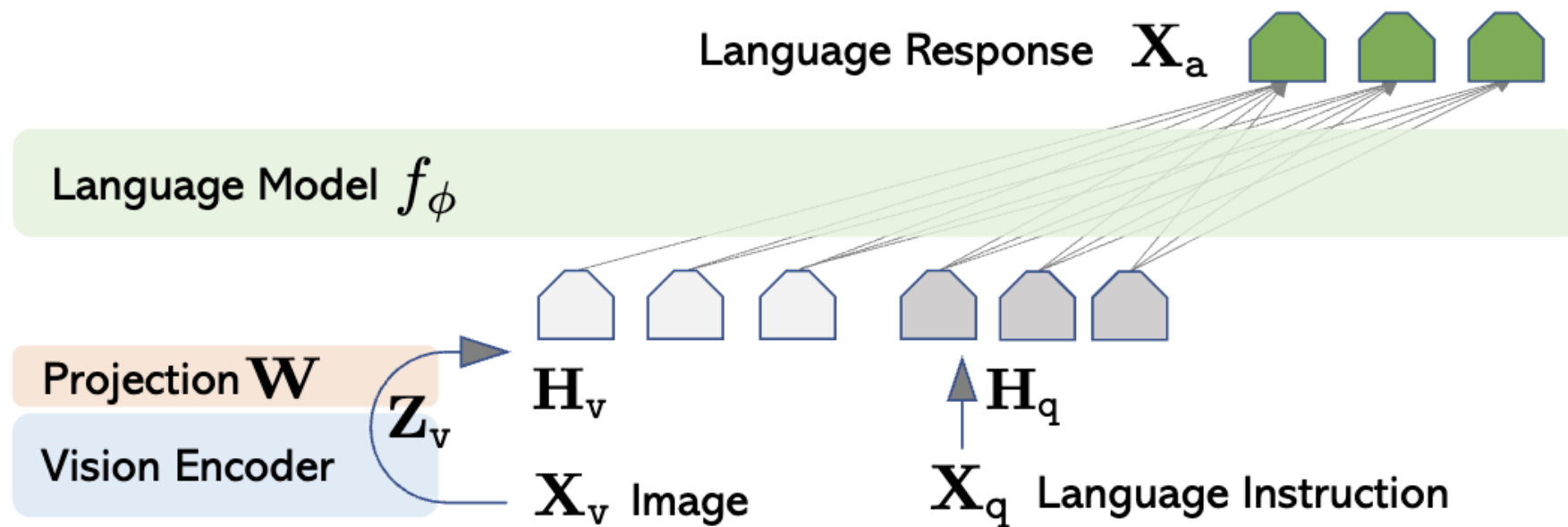
What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

LLaVA: Visual Instruction Tuning



GPT-Assisted Visual Instruction Data Generation

Context type 1: Captions

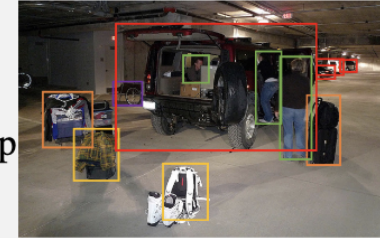
A group of people standing outside of a black vehicle with various luggage.

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], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

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

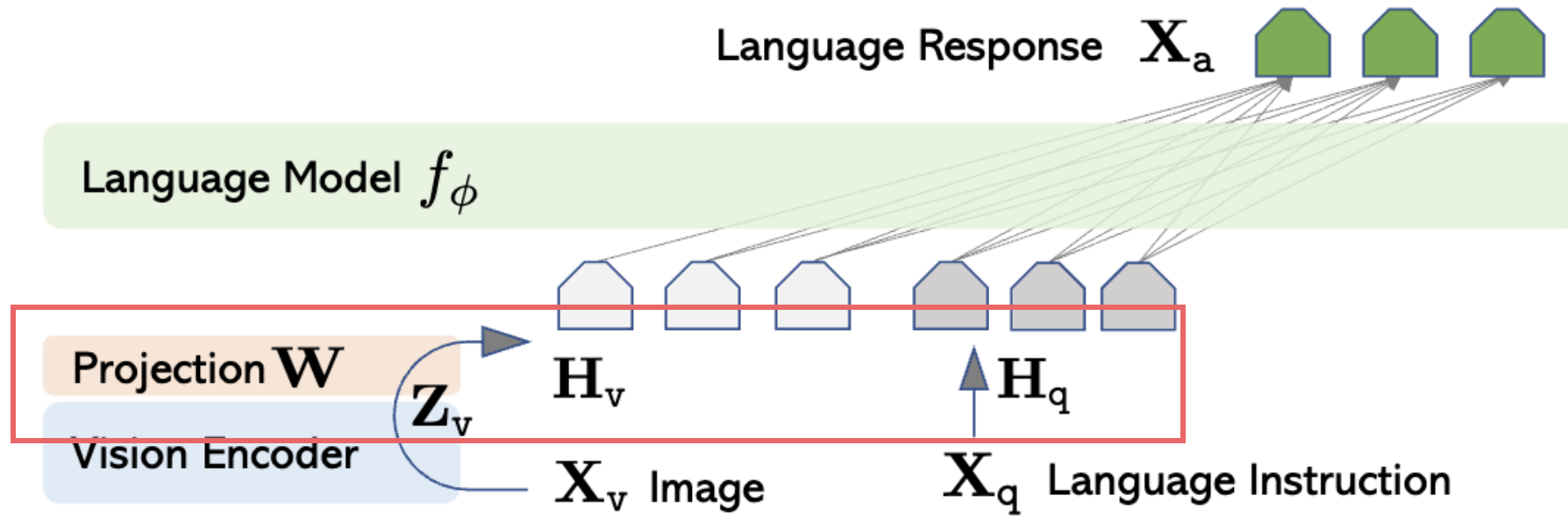
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their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

Response type 3: complex reasoning

Question: What challenges do these people face?

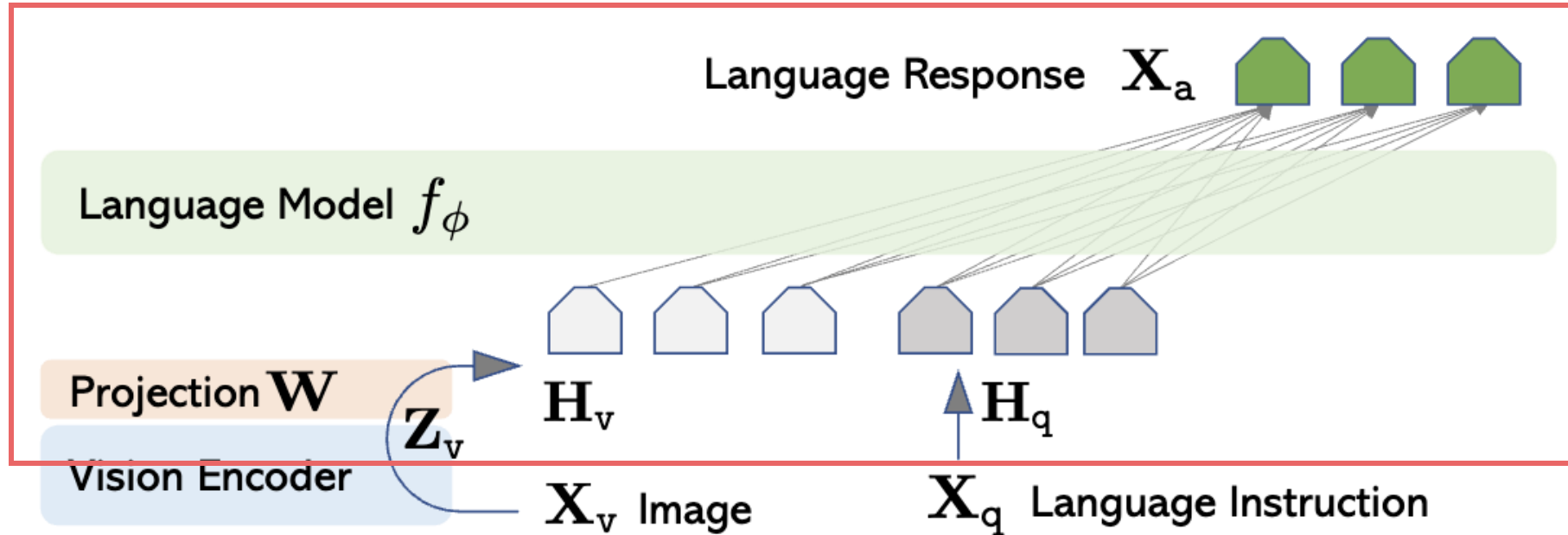
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed,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

Stage 1: Pre-training for Feature Alignment



Train with Image-Text Pairs

Stage 2: Fine-tuning End-to-End



Visual Chat (Visual Instruction Data) and Science QA

Examples

Visual input example, Extreme Ironing:



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

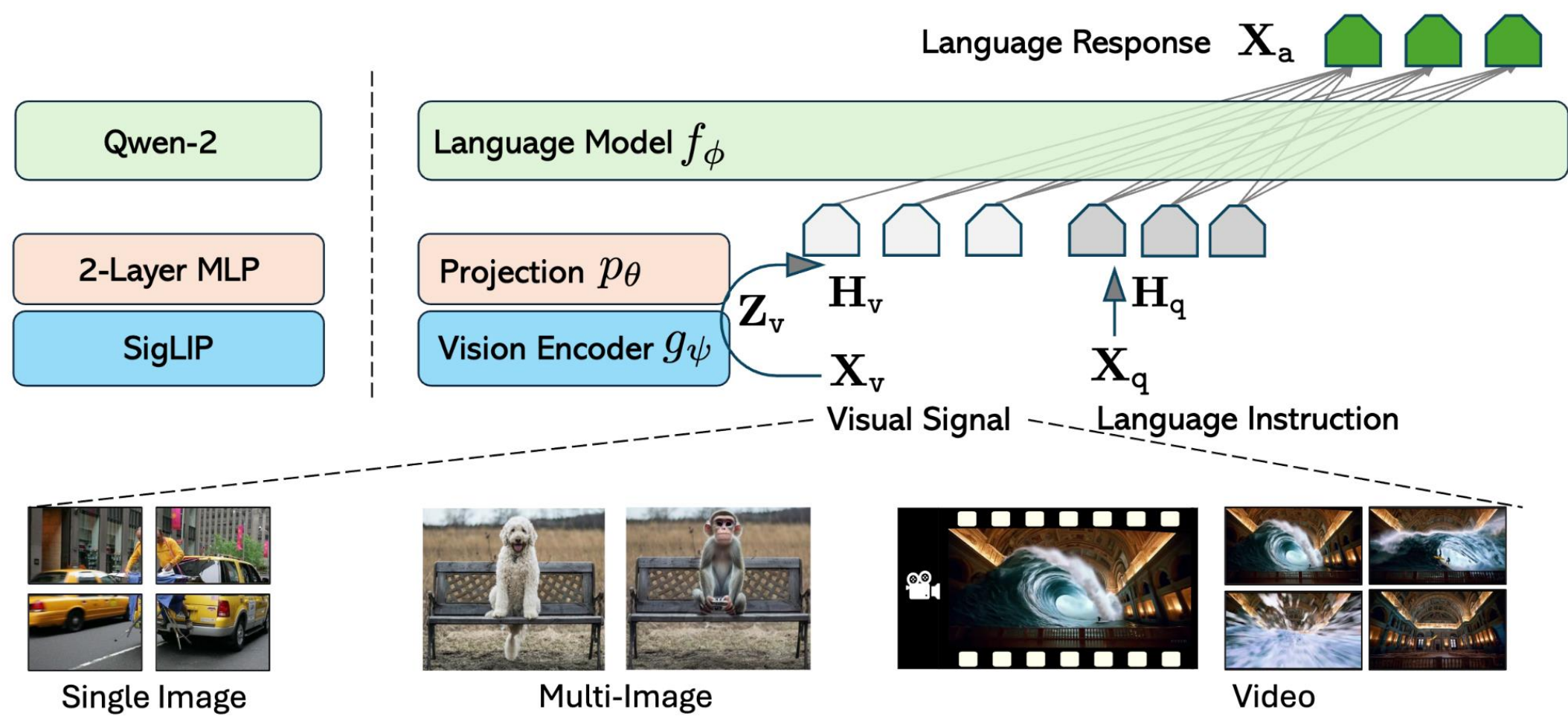
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

User
LLaVA



What is unusual about this image?
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.
[Start a new conversation, and clear the history]
What’s happening in the scene?
The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention **due to his unconventional choice of ironing his clothes on top of a moving car**. The city street around him is bustling with activity, adding to the unique nature of the scene.

User
LLaVA


LLaVA-OneVision



LLaVA-OneVision

 Single-Image	     ... N Crops	$(1 + 9) * 729 = 7290$ Tokens
$729 + N * 729$ Tokens		
 Multi-Image	  ... N Images	$12 * 729 = 8748$ Tokens
$N * 729$ Tokens		
 Video	    ... N Frames	$32 * 196 = 6272$ Tokens
$N * 196$ Tokens		
Example on Token Strategy		Max Tokens


LLaVA-NeXT-Interleave

 **Spot the Difference**


User:
Tell the detailed differences of these two images:




LLaVA-NeXT-M3:
1. In the first image, the player on the top right is wearing a red uniform, but is wearing a blue uniform in the second image.
2. In the first image, the player on the top left has the number "9", but changes to "6" in the second image ...

 **Visual Story Telling**

User:
Complete the story based on the given images and plots:




We had an exciting day at the fair!



The lights lit up the night and the rides made us all dizzy.



The arcade games had the funniest stuffed monkeys as prizes.



LLaVA-NeXT-M3:
We threw a million darts trying to win one!

 **Image Edit Instruction**

User:
Provide the image edit instruction that can transform the source image to the target image:



Source image:



Target image:

LLaVA-NeXT-M3:
Replace the soccer ball with a giant sunflower that the puppy is holding in the same manner as the soccer ball.

 **Social Media Post**

User:
Help me write a Twitter post providing the following images:



LLaVA-NeXT-M3:
Wandering through the enchanting streets of Milan, where the vibrant pulse of the city is as captivating as its historical architecture. From the mesmerizing ... the gentle glow of candles offers a moment of serenity in this bustling metropolis.
#MilanDiaries #TravelMoments #CityLights


 **Low-level Comparison**

User:
Compared to the first image, how is the sharpness of the second image, sharper, blurrier, or about the same?







LLaVA-NeXT-M3: Sharper

 **Video Detailed Description**


User:
Given me a detailed description of the video:



LLaVA-NeXT-M3:
The video depicts an interior scene, specifically a hallway. The flooring is wooden with a light brown hue and a polished finish. There are two doors visible in the video, both with dark frames and glass panels. The door ...

 **3D Grounding**

User:
Where is the bathtub located?



LLaVA-NeXT-M3:
Behind shower curtain inside bathroom

Why Alignment?

- Language modeling \neq assisting users

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.

Why Alignment?

- Language modeling \neq assisting users

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.

Why Alignment?

- Continuing writing does not always work

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

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

A: The answer (arabic numerals) is

(Output) 8 **X**

Instruction Tuning

- LLMs have knowledge, but don't always generate the outputs we want
- Training LLMs to following **human instructions**

Annotated task definitions

You will be given two pieces of text... One of them is simpler ...

You are expected to output 'Text one' if the first sentence is simpler.

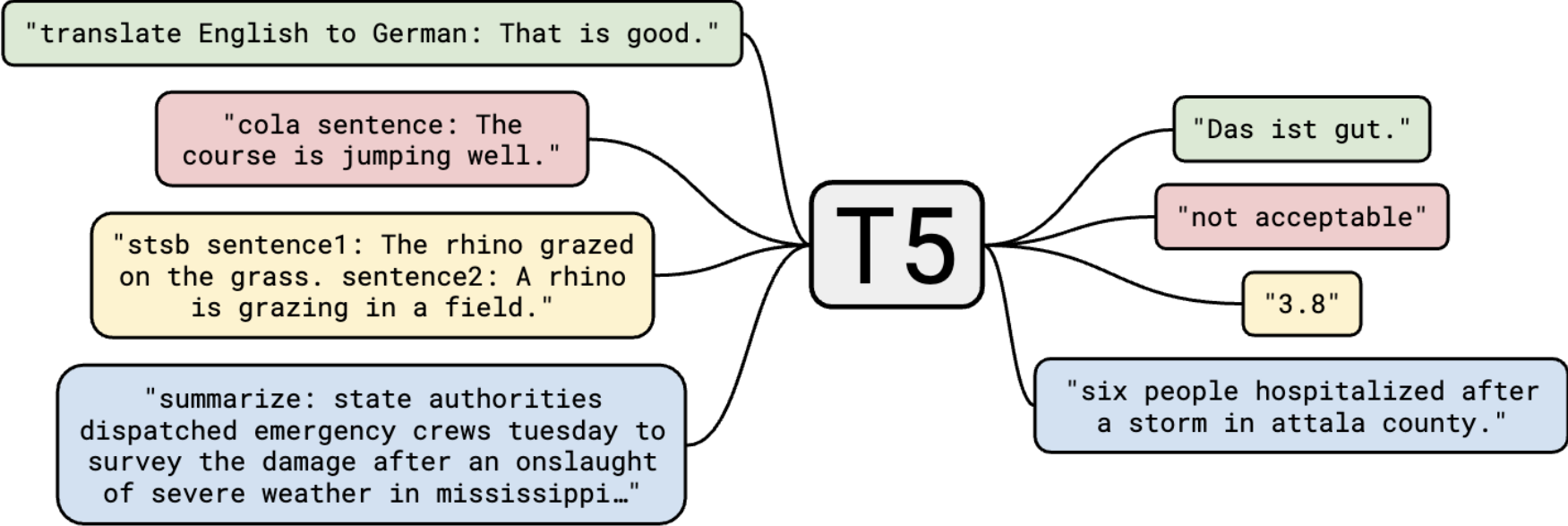
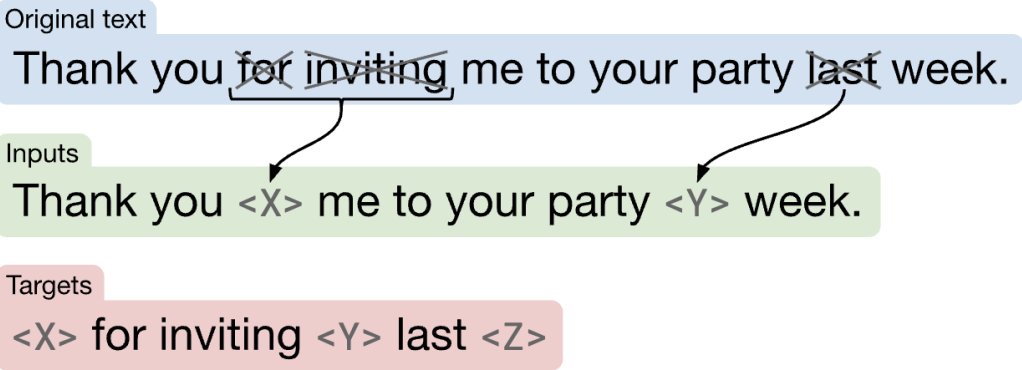
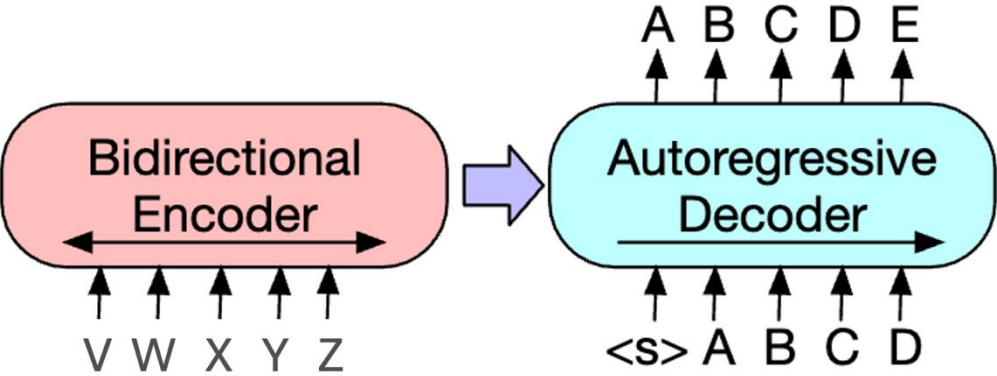
Otherwise output 'Text two'.

Given a sentence with a missing word, pick the answer option that best fills out the missing word in the sentence. Indicate each answer with its index ('a', 'b', 'c', 'd').

Given a document, generate a short title of the document. The title should convey the main idea/event/topic about which the document is being written.

Category	Description
Input Content	Primary description of the task input
<u>Additional Input Content</u>	Additional details on task input
Action Content	Action to perform for task
Input Mention	Mentions of input within action content
Output Content	Primary description of task output
<u>Additional Output Content</u>	Additional details on task output
Label List	Task output labels (classification only)
Label Definition	Task Label definitions (classification only)

Recap: T5



Instruction Tuning

- Convert existing tasks to (input, output) format
- Create many prompts and collect human answers

Annotated task definitions

You will be given two pieces of text... One of them is simpler ...

You are expected to output 'Text one' if the first sentence is simpler.

Otherwise output 'Text two'.

Given a sentence with a missing word, pick the answer option that best fills out the missing word in the sentence. Indicate each answer with its index ('a', 'b', 'c', 'd').

Given a document, generate a short title of the document. The title should convey the main idea/event/topic about which the document is being written.

Category	Description
Input Content	Primary description of the task input
Additional Input Content	Additional details on task input
Action Content	Action to perform for task
Input Mention	Mentions of input within action content
Output Content	Primary description of task output
Additional Output Content	Additional details on task output
Label List	Task output labels (classification only)
Label Definition	Task Label definitions (classification only)

Sidenote: Why Decoder-Only Instead of Encoder-Decoder?

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

Binge ... on | - | and | of | is

Binge **drinking** ... is | and | had | in | was

Binge drinking **may** ... be | also | have | not | increase

Binge drinking may **not** ... be | have | cause | always | help

Binge drinking may not **necessarily** ... be | lead | cause | results | have

Binge drinking may not necessarily **kill** ... you | the | a | people | your

Binge drinking may not necessarily kill **or** ... even | injure | kill | cause | prevent

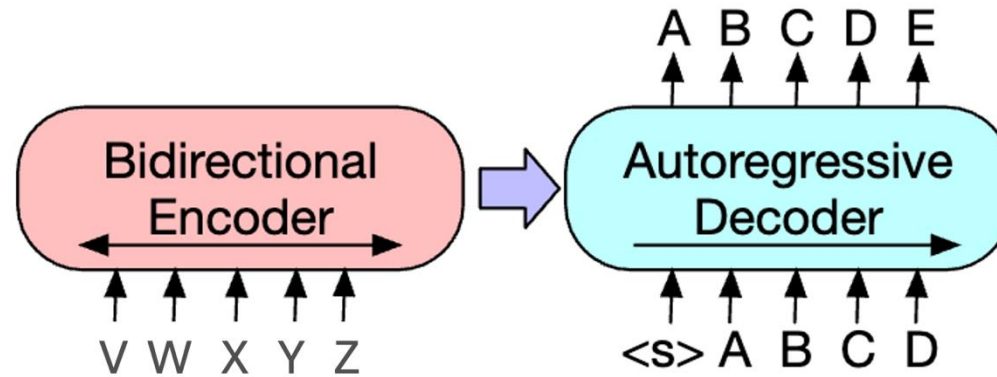
Binge drinking may not necessarily kill or **even** ... kill | prevent | cause | reduce | injure

Binge drinking may not necessarily kill or even **damage** ... your | the | a | you | someone

Binge drinking may not necessarily kill or even damage **brain** ... cells | functions | tissue | neurons

Binge drinking may not necessarily kill or even damage brain **cells**, ... some | it | the | is | long

Sidenote: Why Decoder-Only Instead of Encoder-Decoder?

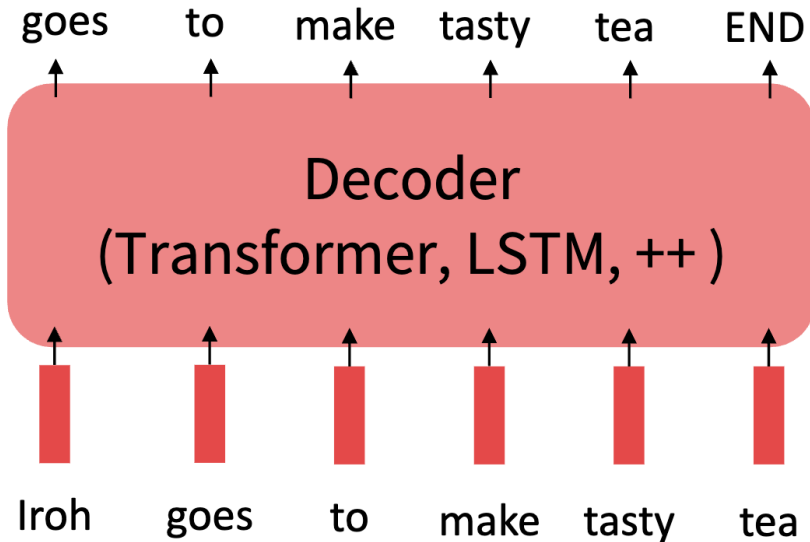


Binge ... on | - | and | of | is
Binge **drinking** ... is | and | had | in | was
Binge drinking **may** ... be | also | have | not | increase
Binge drinking may **not** ... be | have | cause | always | help
Binge drinking may not **necessarily** ... be | lead | cause | results | have
Binge drinking may not necessarily **kill** ... you | the | a | people | your
Binge drinking may not necessarily kill **or** ... even | injure | kill | cause | prevent
Binge drinking may not necessarily kill or **even** ... kill | prevent | cause | reduce | injure
Binge drinking may not necessarily kill or even **damage** ... your | the | a | you | someone
Binge drinking may not necessarily kill or even damage **brain** ... cells | functions | tissue | neurons
Binge drinking may not necessarily kill or even damage brain **cells**, ... some | it | the | is | long

Scaling Up Instruction Tuning

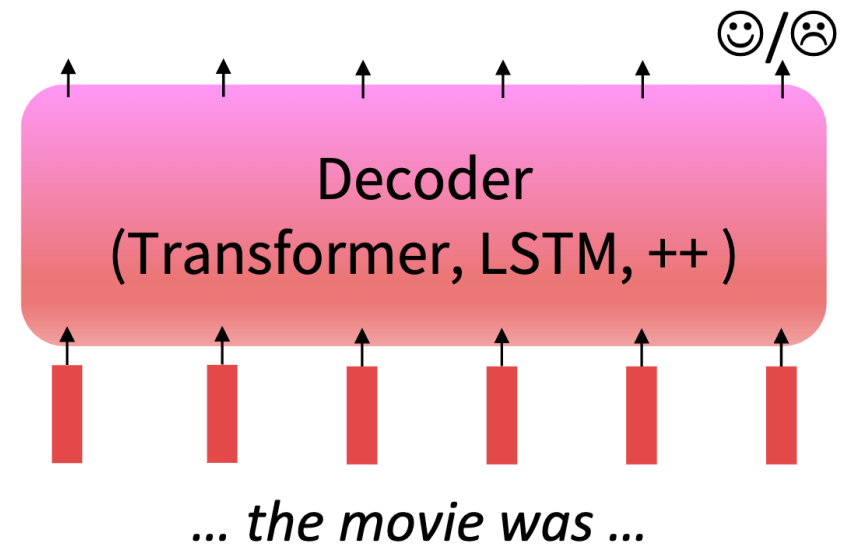
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



Step 2: Finetune (on **many tasks**)

Not many labels; adapt to the tasks!



Instruction Tuning → Instruction Pre-Training

- Instruction fine-tuning for many tasks



Instruction Tuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✗ (doesn't answer question)

Instruction Tuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

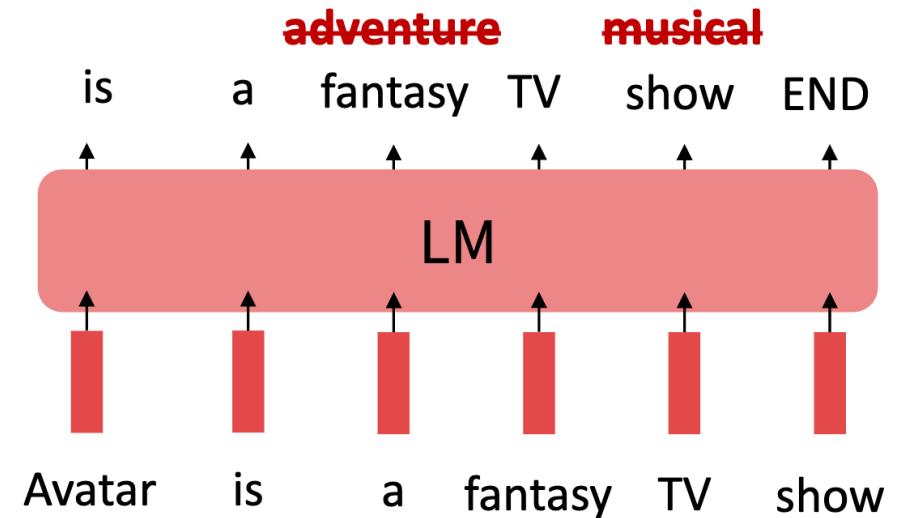
After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

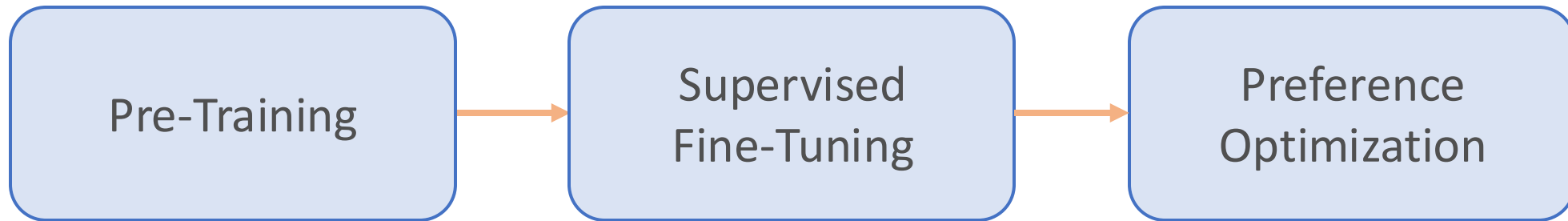
Limitations of Instruction Fine-Tuning

- It is expensive to collect ground-truth data for tasks
- Open-ended creative generation have no right answer
 - E.g., write me a story about a dog and her pet grasshopper
- language modeling penalizes all token-level mistakes equally, but some errors are worse than others

Even with instruction finetuning, there is still a mismatch between the LM objective and “satisfying human preferences”!



Alignment Pipeline



Reinforcement Learning from Human Feedback (RLHF)



Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*

OpenAI

Human Feedback

- Human reward

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco

...
overturn unstable
objects.

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

$$s_1 \\ R(s_1) = 8.0$$

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

$$s_2 \\ R(s_2) = 1.2$$

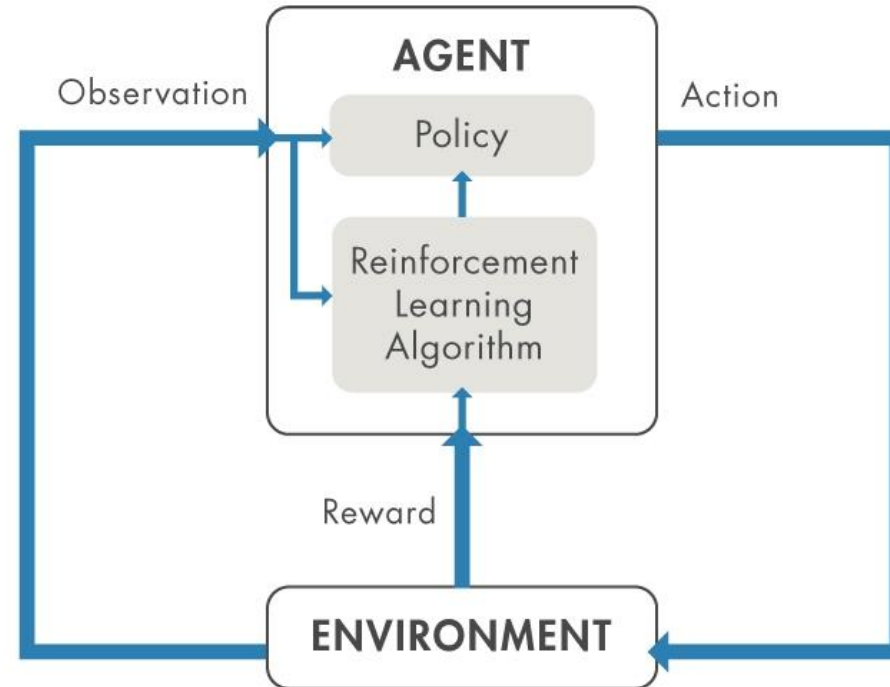
Goal: maximize the expected reward of samples from our LM

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$$

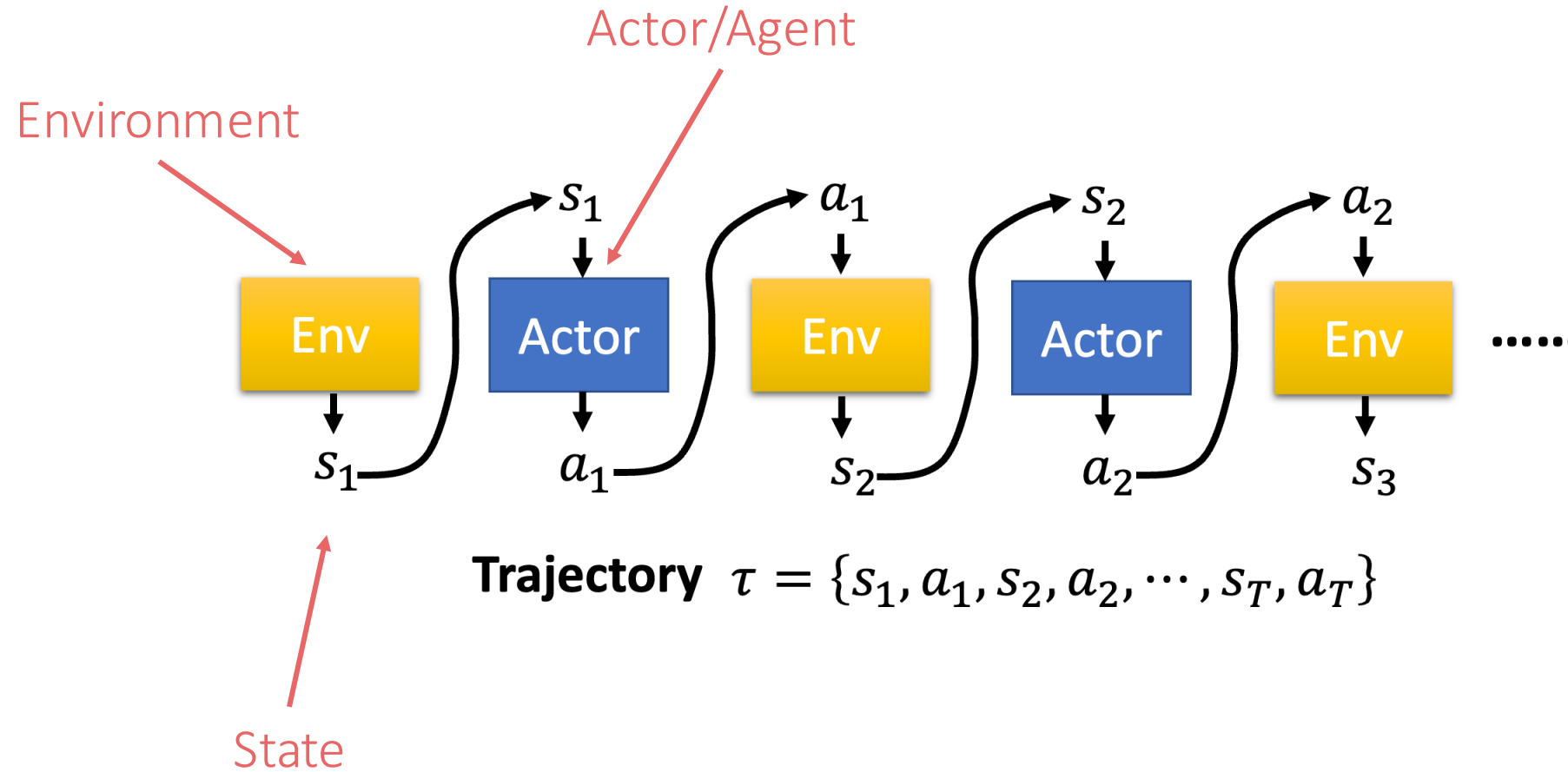
Reinforcement Learning from Human Preferences

How do we change the LM parameters θ to maximize this?

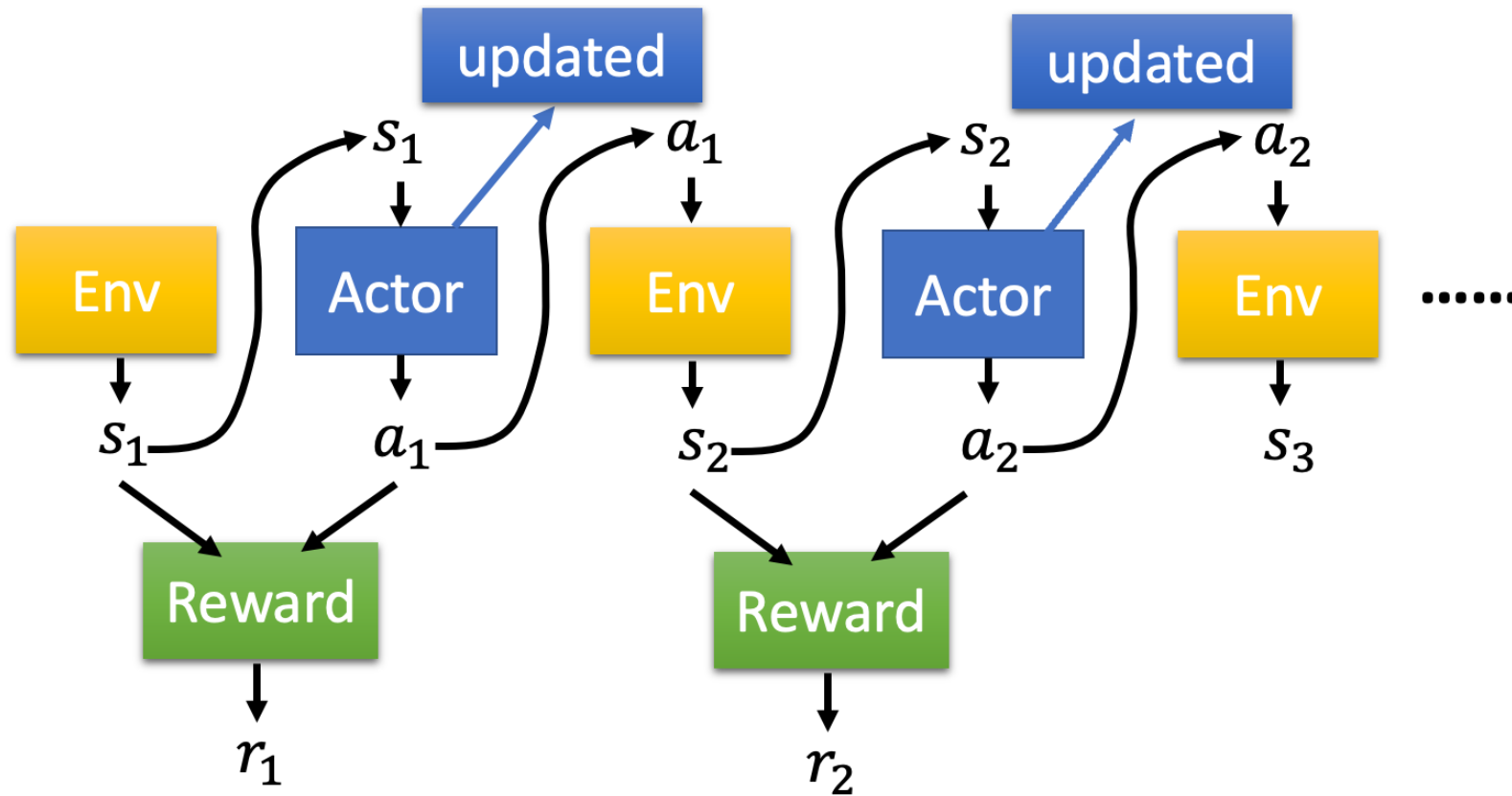
$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$$



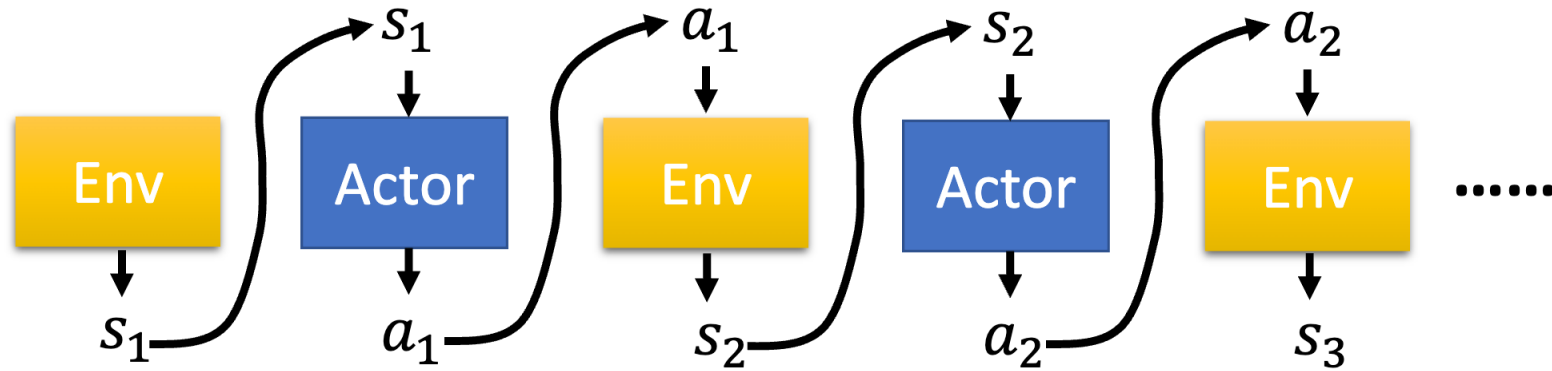
Reinforcement Learning



Reinforcement Learning



Reinforcement Learning



Trajectory $\tau = \{s_1, a_1, s_2, a_2, \dots, s_T, a_T\}$

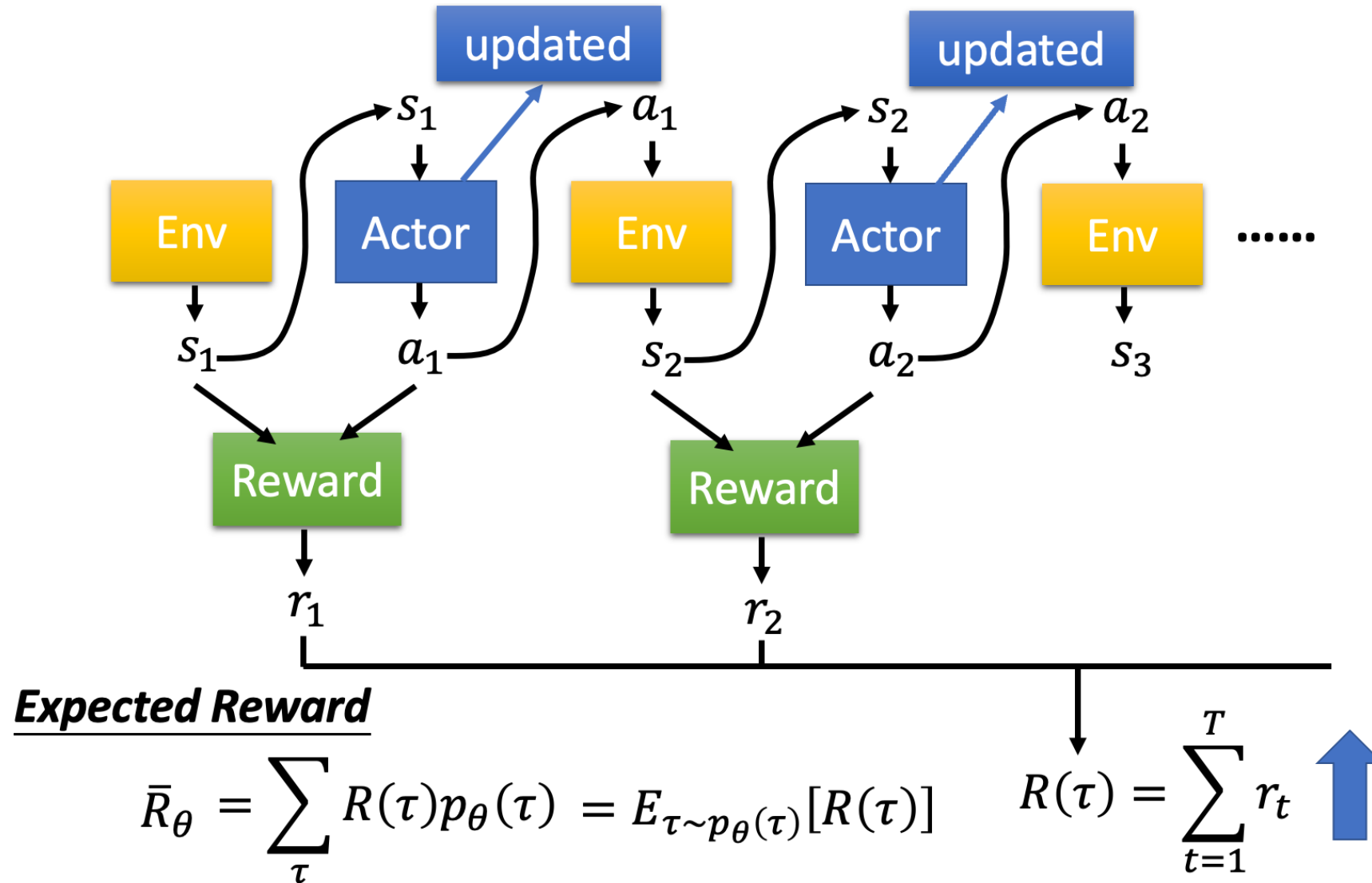
$$p_{\theta}(\tau)$$

$$= p(s_1)p_{\theta}(a_1|s_1)p(s_2|s_1, a_1)p_{\theta}(a_2|s_2)p(s_3|s_2, a_2) \dots$$

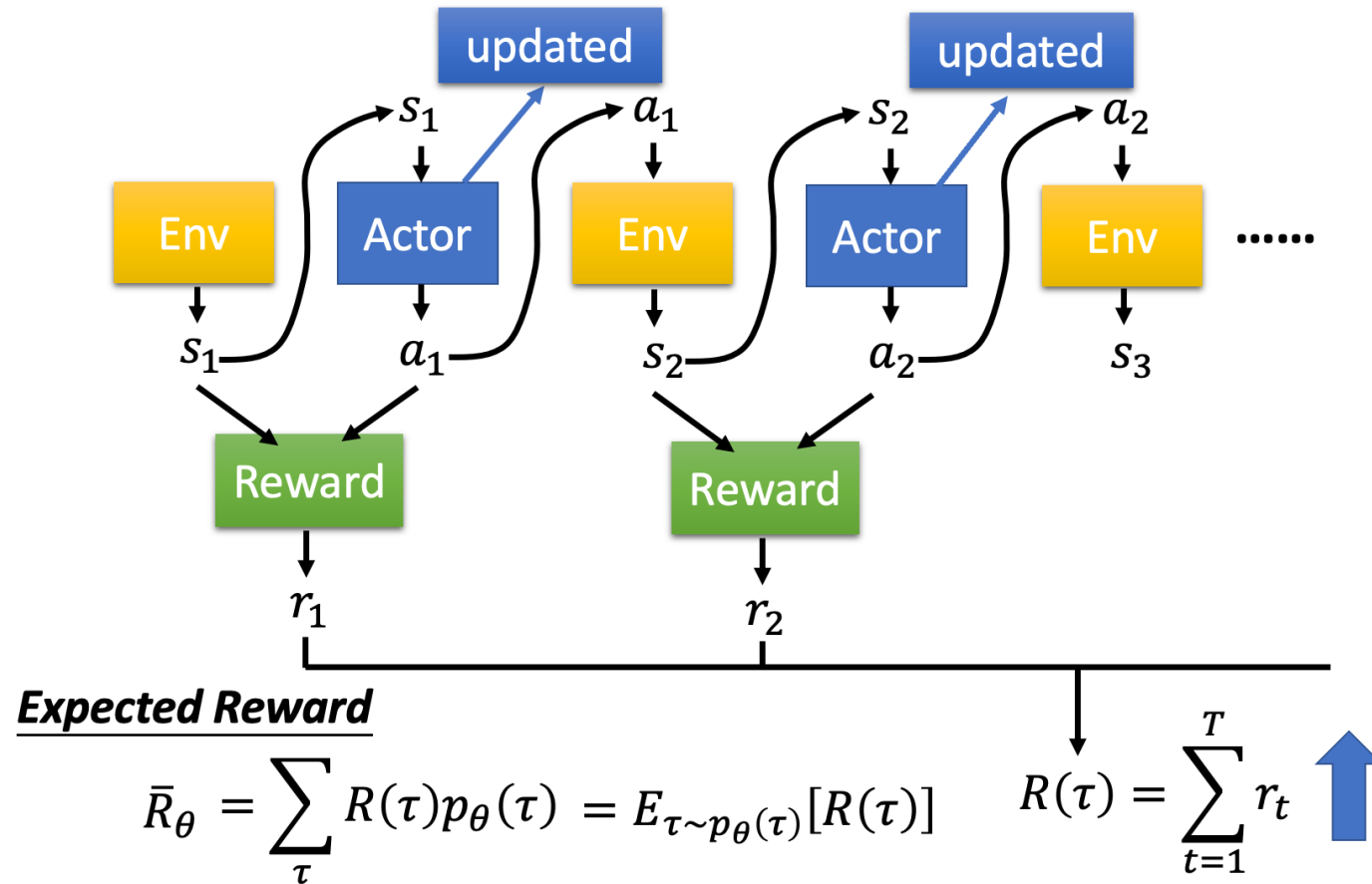
$$= p(s_1) \prod_{t=1}^T p_{\theta}(a_t|s_t)p(s_{t+1}|s_t, a_t)$$

https://blog.csdn.net/qq_30615903

Reinforcement Learning



Reinforcement Learning



Solutions

- Q-Learning
- Policy Gradient
- Actor-Critic
- ...

Optimizing for Human Preferences

How do we change the LM parameters θ to maximize this?

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$$

Gradient **Ascent**

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)}[R(\hat{s})]$$

Policy Gradient Methods in Reinforcement Learning
(REINFORCE) [Williams, 1992]

Policy Gradient/REINFORCE

Gradient Ascent

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)} [R(\hat{s})]$$

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \nabla_{\theta} \sum_s R(s) p_{\theta}(s) = \sum_s R(s) \nabla_{\theta} p_{\theta}(s)$$

Log-Derivative Trick

$$\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s) \quad \Rightarrow \quad \nabla_{\theta} p_{\theta}(s) = \nabla_{\theta} \log p_{\theta}(s) p_{\theta}(s)$$

Policy Gradient/REINFORCE

$$\begin{aligned}\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})] &= \sum_s R(s) \nabla_{\theta} p_{\theta}(s) = \sum_s p_{\theta}(s) R(s) \nabla_{\theta} \log p_{\theta}(s) \\ &= \mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})]\end{aligned}$$

We can approximate this objective with Monte Carlo samples

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})] = \mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})] \approx \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$$

Policy Gradient/REINFORCE

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

Diagram illustrating the REINFORCE update rule with annotations:

- If R is +++** (Green text, arrow pointing to $R(s_i)$)
- Take gradient steps to maximize $p_{\theta}(s_i)$** (Green text, arrow pointing to $\nabla_{\theta_t} \log p_{\theta_t}(s_i)$)
- If R is ---** (Red text, arrow pointing to $R(s_i)$)
- Take steps to minimize $p_{\theta}(s_i)$** (Red text, arrow pointing to $\nabla_{\theta_t} \log p_{\theta_t}(s_i)$)

We **reinforce** good actions, increasing the chance they happen again

Proximal Policy Optimization (PPO)

- New parameters θ' cannot be very different from old parameters θ

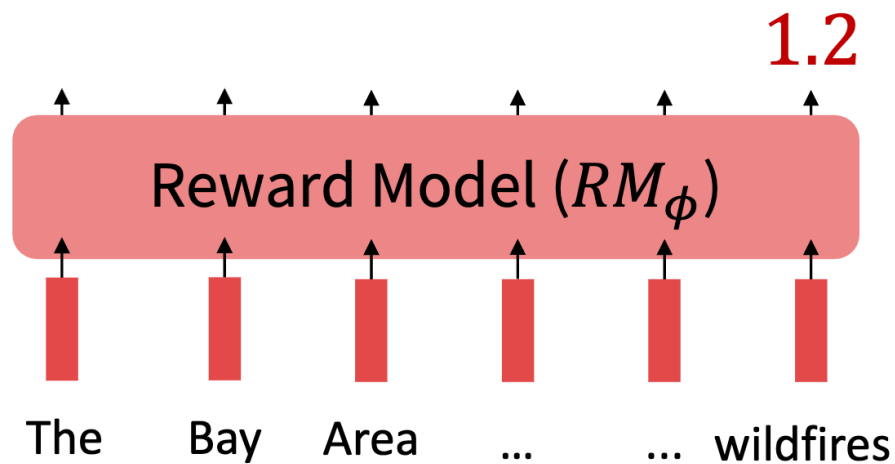
$$J_{PPO}^{\theta'}(\theta) = J^{\theta'}(\theta) - \beta KL(\theta, \theta')$$



Regularization

How to Model Human Preferences?

- Now for any reward function R , we can train our language model to maximize expected reward
- Problem 1: human-in-the-loop is expensive
 - Solution: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem
 - Train a reward model (RM) from an annotated dataset



How to Model Human Preferences?

- Now for any reward function R , we can train our language model to maximize expected reward
- Problem 2: human judgments are noisy and miscalibrated
 - Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

S_1

>

A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.

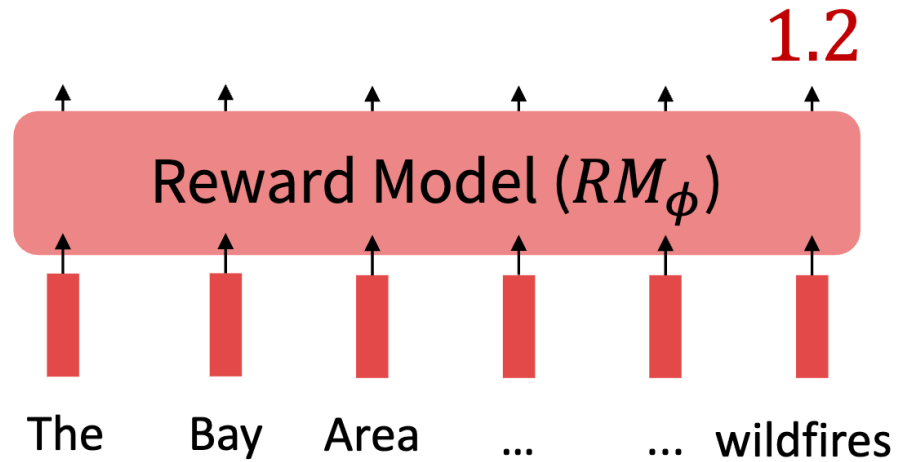
S_3

>

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

S_2

Training A Reward Model



Bradley-Terry [1952] paired comparison model

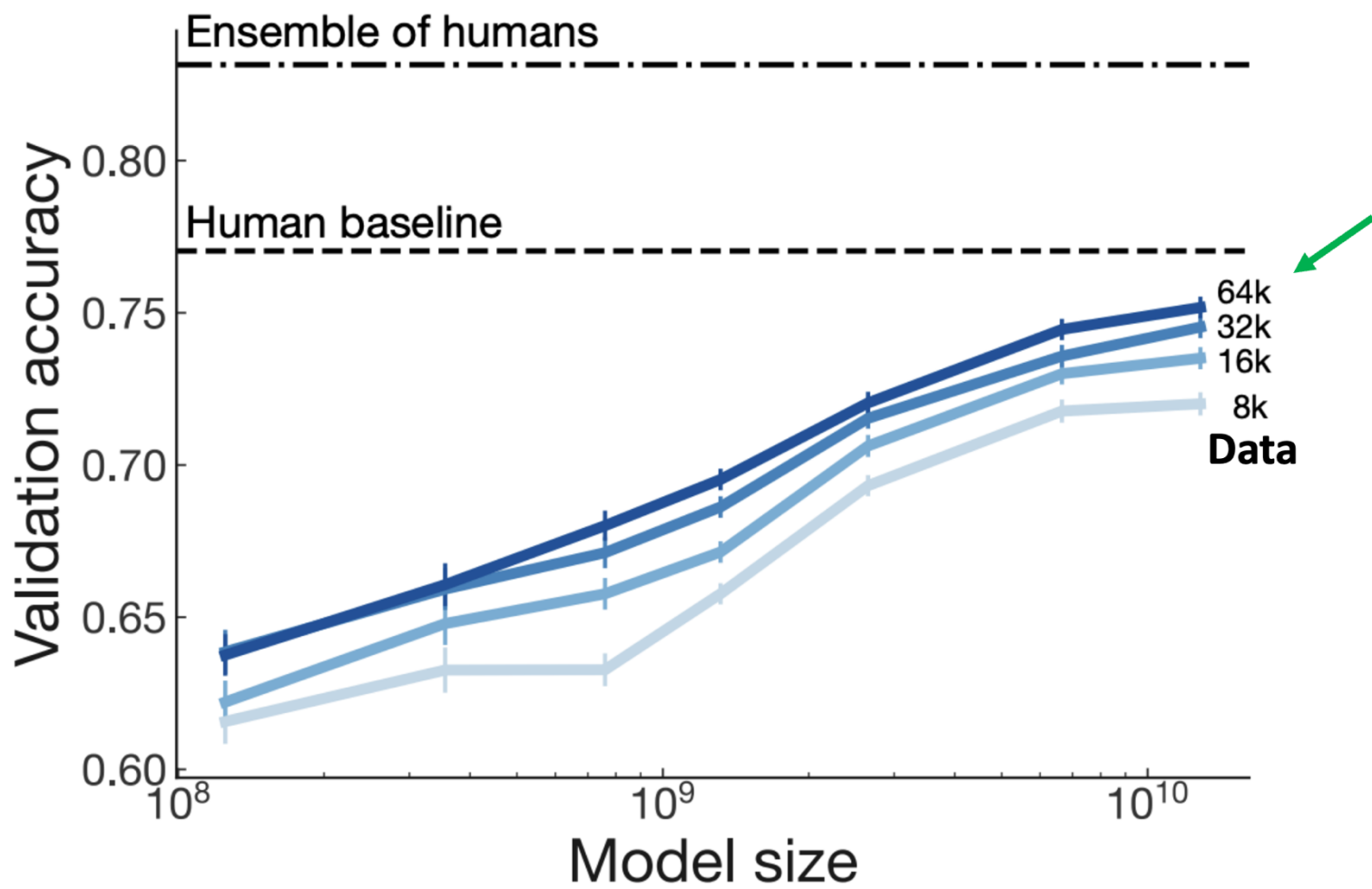
$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))]$$

“winning”
sample

“losing”
sample

s^w should score
higher than s^l

Reward Model vs. Real Human Feedback



Large enough RM
trained on enough
data approaching
single human perf

[Stiennon et al., 2020]

RLHF: Putting Everything All Together

- We have the following:
 - A pretrained (possibly instruction-finetuned) LM $p^{PT}(y | x)$
 - A reward model $RM_{\phi}(x, y)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
- Now to do RLHF:
 - Copy the model $p_{\theta}^{RL}(y | x)$, with parameters θ we would like to optimize
 - We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y} | x)} [RM_{\phi}(x, \hat{y})]$$

RLHF: Putting Everything All Together

- We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y} | x)} [RM_{\phi}(x, \hat{y})]$$

- Do you see any problems?
 - Learned rewards are imperfect; this quantity can be imperfectly optimized
- Add a penalty for drifting too far from the initialization:

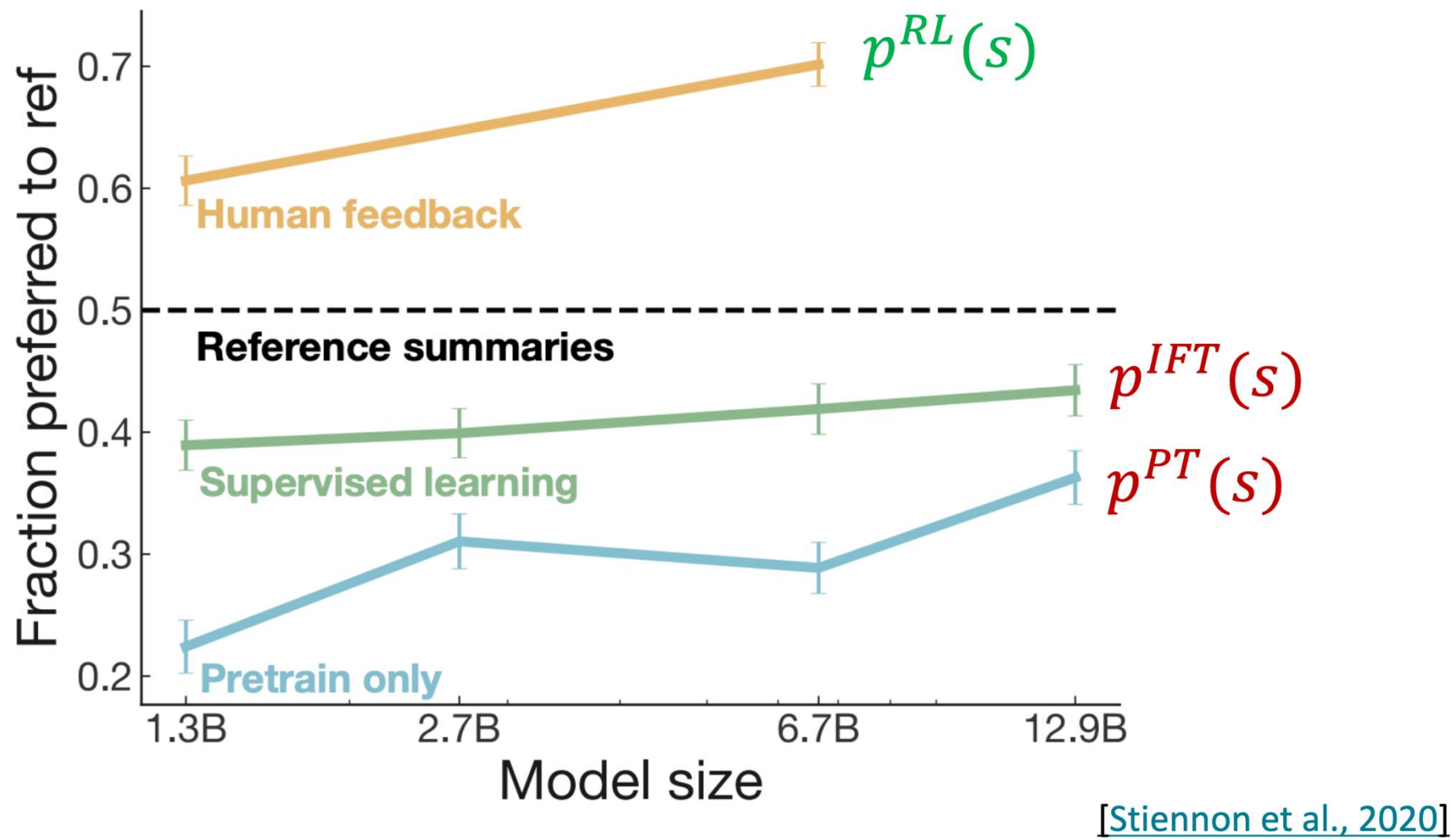
$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y} | x)} [RM_{\phi}(x, \hat{y}) - \underbrace{\beta \log \left(\frac{p_{\theta}^{RL}(\hat{y} | x)}{p^{PT}(\hat{y} | x)} \right)}]$$

Pay a price when

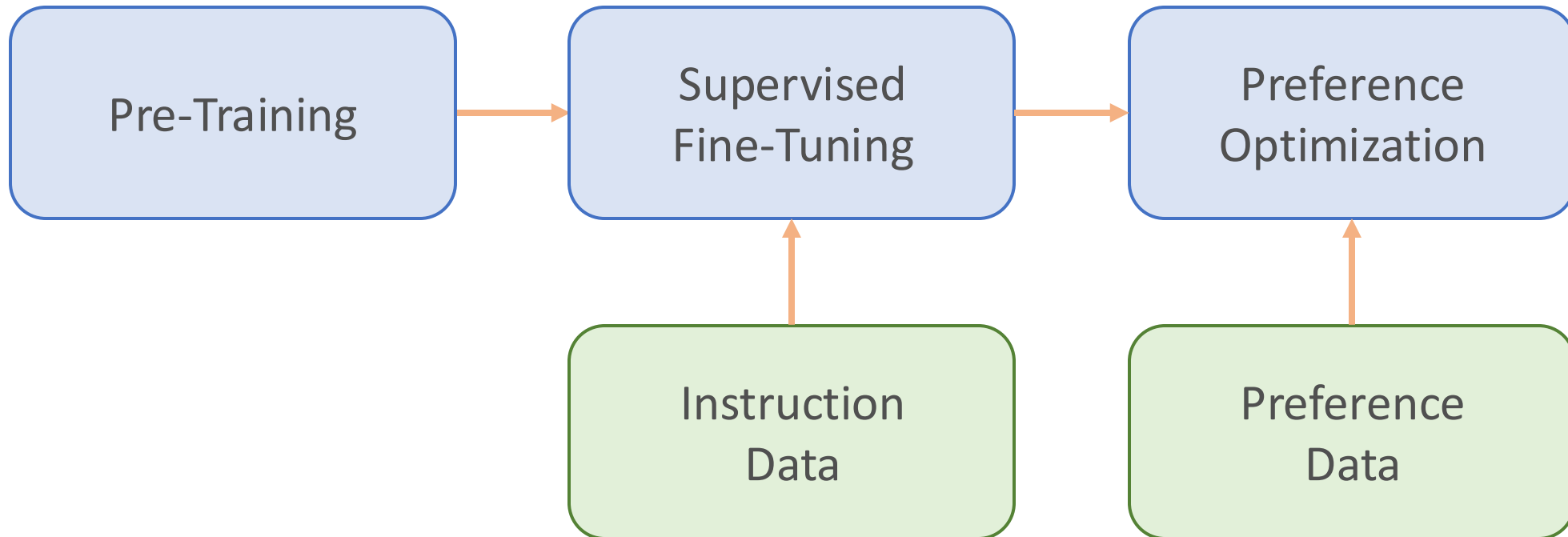
$$p_{\theta}^{RL}(\hat{y} | x) > p^{PT}(\hat{y} | x)$$

This penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL)** divergence between $p_{\theta}^{RL}(\hat{y} | x)$ and $p^{PT}(\hat{y} | x)$.

RLHF vs. Supervised Fine-Tuning



Alignment Pipeline

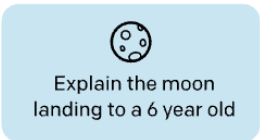


InstructGPT

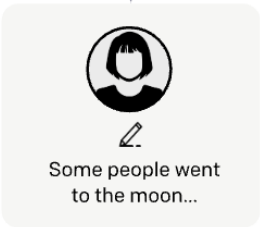
Step 1

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

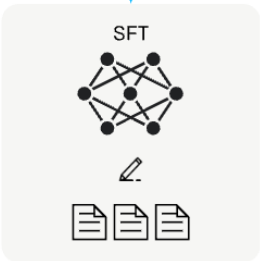
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



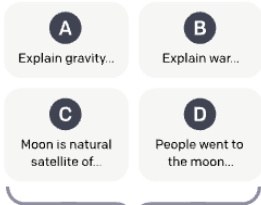
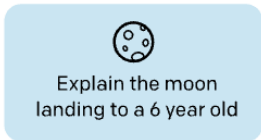
This data is used
to fine-tune GPT-3
with supervised
learning.



Step 2

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

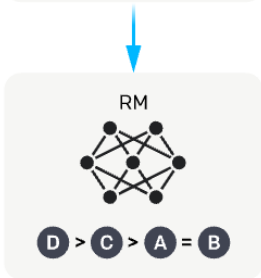
A prompt and
several model
outputs are
sampled.



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.

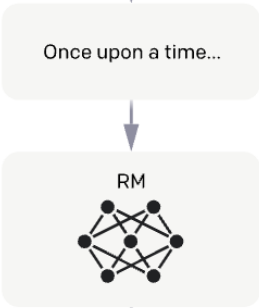


The policy
generates an output.

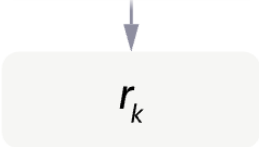


Once upon a time...

The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.



ChatGPT: Instruction Fine-tuning + RLHF for Dialog Agents

ChatGPT: Optimizing Language Models for Dialogue

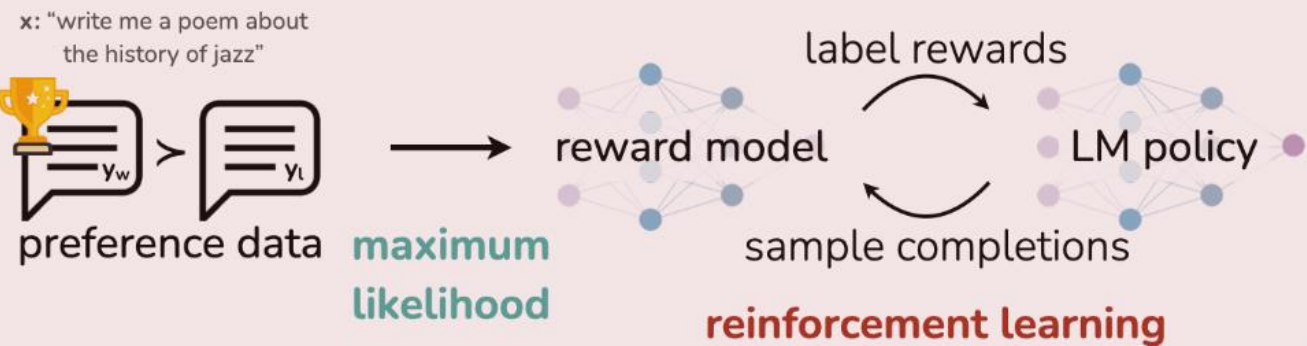
Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

Direct Preference Optimization (DPO)

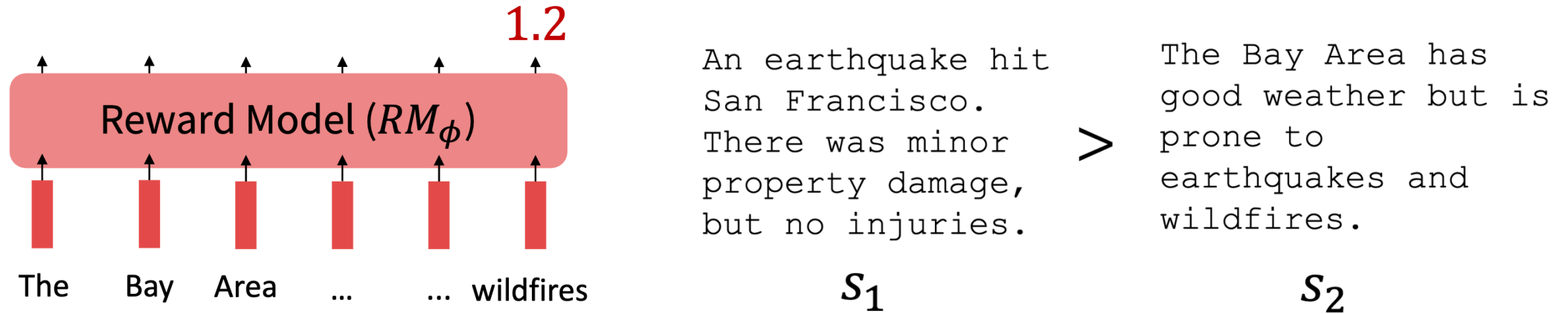
Reinforcement Learning from Human Feedback (RLHF)



Direct Preference Optimization (DPO)



RLHF: Proximal Policy Optimization (PPO)



$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))]$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay close to reference model)

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(y|x)} [r(x, y)] - \beta \mathbb{D}_{\text{KL}}(\pi(\cdot | x) || \pi_{\text{ref}}(\cdot | x))$$

Maximize reward

Keep similar behavior

$$\begin{aligned} & \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta \mathbb{D}_{\text{KL}}[\pi(y|x) || \pi_{\text{ref}}(y|x)] \\ &= \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[r(x, y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) \right] \\ &= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} - \log Z(x) \right] \\ & Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \end{aligned}$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay close to reference model)

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(y|x)} [r(x, y)] - \beta \mathbb{D}_{\text{KL}}(\pi(\cdot | x) || \pi_{\text{ref}}(\cdot | x))$$

Maximize reward

Keep similar behavior

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right) \quad \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right)} - \log Z(x) \right]$$

$$= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)} \right] - \log Z(x) \right]$$

$$= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} [\mathbb{D}_{\text{KL}}(\pi(y|x) || \pi^*(y|x)) - \log Z(x)]$$

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right)$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay close to reference model)

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(y|x)} [r(x, y)] - \beta \mathbb{D}_{\text{KL}}(\pi(\cdot | x) \| \pi_{\text{ref}}(\cdot | x))$$

Maximize reward

Keep similar behavior

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

$$\pi^*(y | x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y | x) \exp \left(\frac{1}{\beta} r(x, y) \right)$$

with $Z(x) = \sum_y \pi_{\text{ref}}(y | x) \exp \left(\frac{1}{\beta} r(x, y) \right)$

Note **intractable sum** over possible responses; can't immediately use this

Rearrange

(write **any reward function** as function of **optimal policy**)

$$r(x, y) = \underbrace{\beta \log \frac{\pi^*(y | x)}{\pi_{\text{ref}}(y | x)}}_{\text{some parameterization of a reward function}} + \beta \log Z(x)$$

Ratio is **positive** if policy likes response more than reference model, **negative** if policy likes response less than ref. model

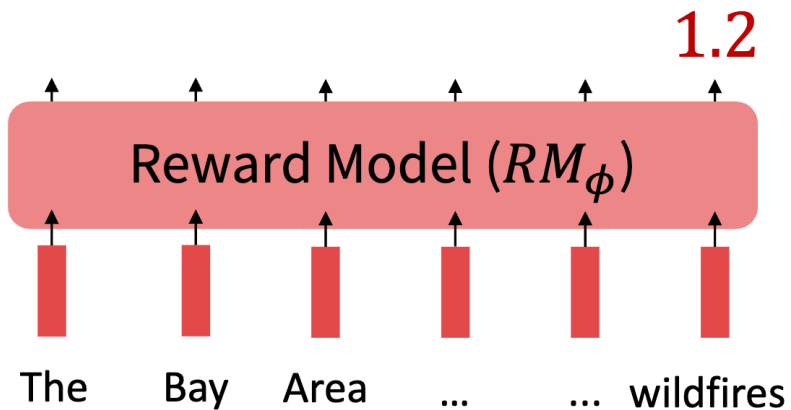
some parameterization of a reward function

Direct Preference Optimization (DPO)

A loss function on reward functions

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r(x, y_w) - r(x, y_l))]$$



An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

s_1

>

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

s_2

Direct Preference Optimization (DPO)

**A loss function on
reward functions**

+

**A transformation
between reward
functions and policies**

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r(x, y_w) - r(x, y_l))]$$

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

Direct Preference Optimization (DPO)

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r(x, y_w) - r(x, y_l))]$$

+

A transformation
between reward
functions and policies

=

A loss function
on policies

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x)$$

Reward of
preferred
response

Reward of
dispreferred
response

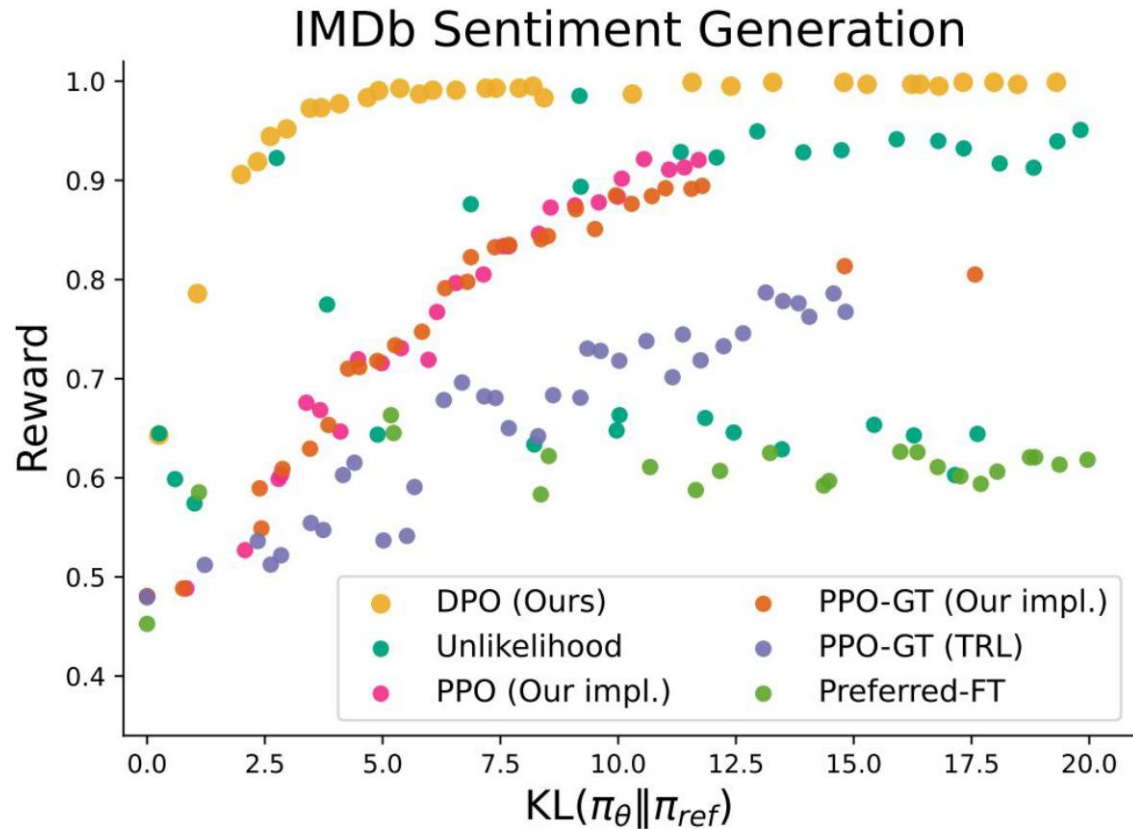
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

Direct Preference Optimization (DPO)

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\underbrace{\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)}}_{\text{Reward of preferred response}} - \underbrace{\beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)}}_{\text{Reward of dispreferred response}} \right) \right]$$

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = \\ -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\underbrace{\sigma(\hat{r}_{\theta}(x, y_l) - \hat{r}_{\theta}(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{\nabla_{\theta} \log \pi(y_w | x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_{\theta} \log \pi(y_l | x)}_{\text{decrease likelihood of } y_l} \right] \right] \end{aligned}$$

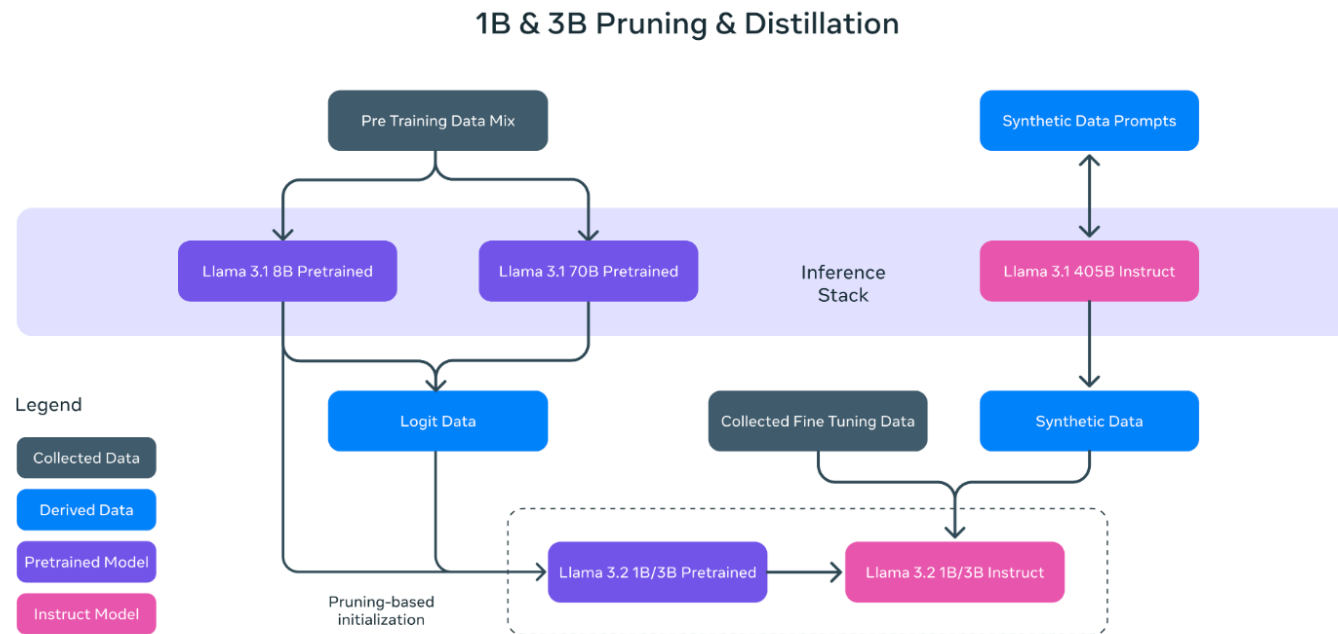
DPO Performance



1. Generate positive IMDB reviews from GPT2-XL
2. Use pre-trained sentiment classifier as Gold RM
3. Create preferences based on Gold RM
4. Optimize with PPO and DPO

Large-Scale DPO Training

Llama 3.2: Revolutionizing edge AI and vision with open, customizable models



In post-training, we use a similar recipe as Llama 3.1 and produce final chat models by doing several rounds of alignment on top of the pre-trained model. Each round involves supervised fine-tuning (SFT), rejection sampling (RS), and direct preference optimization (DPO).

Simple Preference Optimization (SimPO)

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

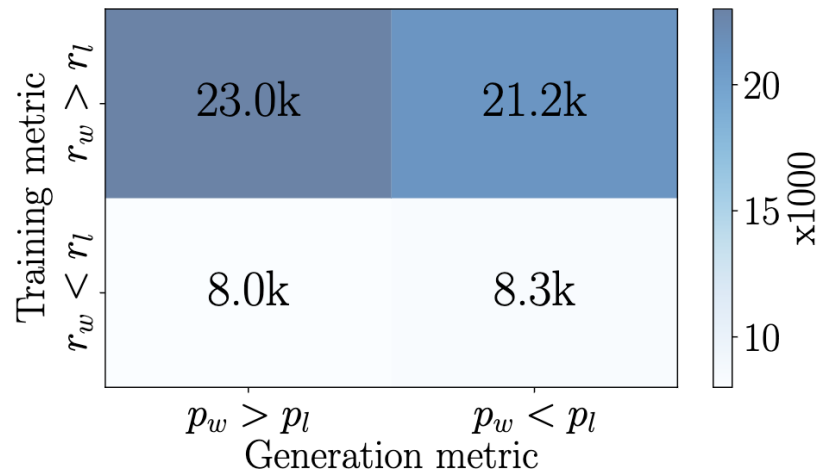
$$\mathcal{L}_{\text{SimPO}}(\pi_{\theta}) = -\mathbb{E} \left[\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_{\theta}(y_w | x) - \frac{\beta}{|y_l|} \log \pi_{\theta}(y_l | x) - \gamma \right) \right]$$

Look Back at DPO

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \underbrace{\frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)}}_{\text{Reward of preferred response}} - \beta \log \underbrace{\frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)}}_{\text{Reward of dispreferred response}} \right) \right]$$

How does reference model affect the behavior?

$$r(x, y_w) > r(x, y_l) \Rightarrow p_{\theta}(y_w | x) > p_{\theta}(y_l | x)?$$



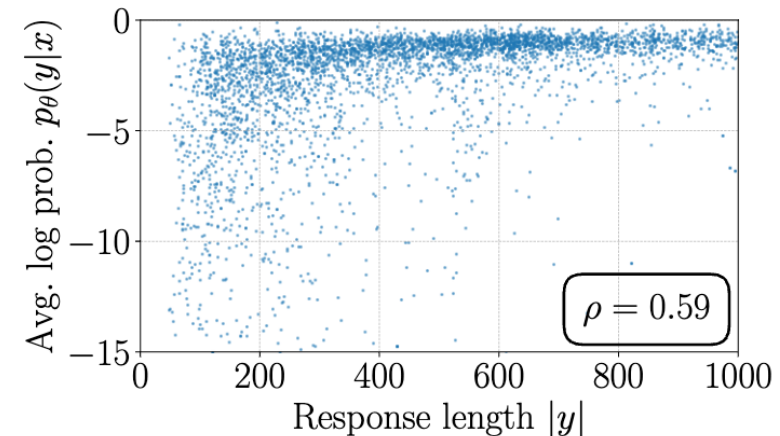
Solution: Reference-Free Reward

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\underbrace{\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)}}_{\text{Reward of preferred response}} - \underbrace{\beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)}}_{\text{Reward of dispreferred response}} \right) \right]$$

$$r(x, y) = \sum_{i=1}^{|y|} \log \pi_{\theta}(y_i | x, y_{<i})$$

Length bias!

The model tends to generate longer sequence to maximize reward



(a) Length correlation (DPO).

Solution: Reference-Free Reward

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\underbrace{\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)}}_{\text{Reward of preferred response}} - \underbrace{\beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)}}_{\text{Reward of dispreferred response}} \right) \right]$$

$$r_{\text{SimPO}}(x, y) = \frac{\beta}{|y|} \log \pi_{\theta}(y | x) = \frac{\beta}{|y|} \sum_{i=1}^{|y|} \log \pi_{\theta}(y_i | x, y_{<i})$$

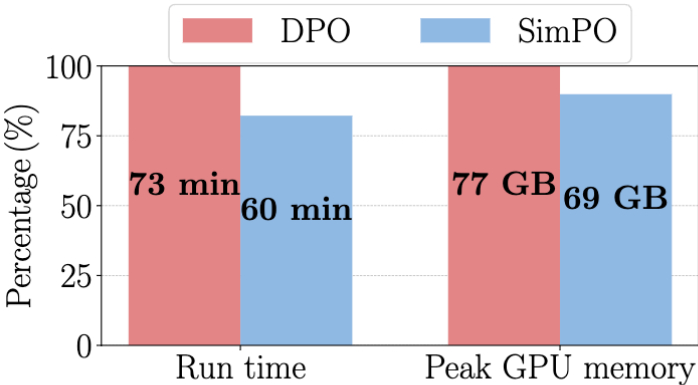
Reward margin

$$\mathcal{L}_{\text{SimPO}}(\pi_{\theta}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_{\theta}(y_w | x) - \frac{\beta}{|y_l|} \log \pi_{\theta}(y_l | x) - \boxed{\gamma} \right) \right]$$

SimPO Performance

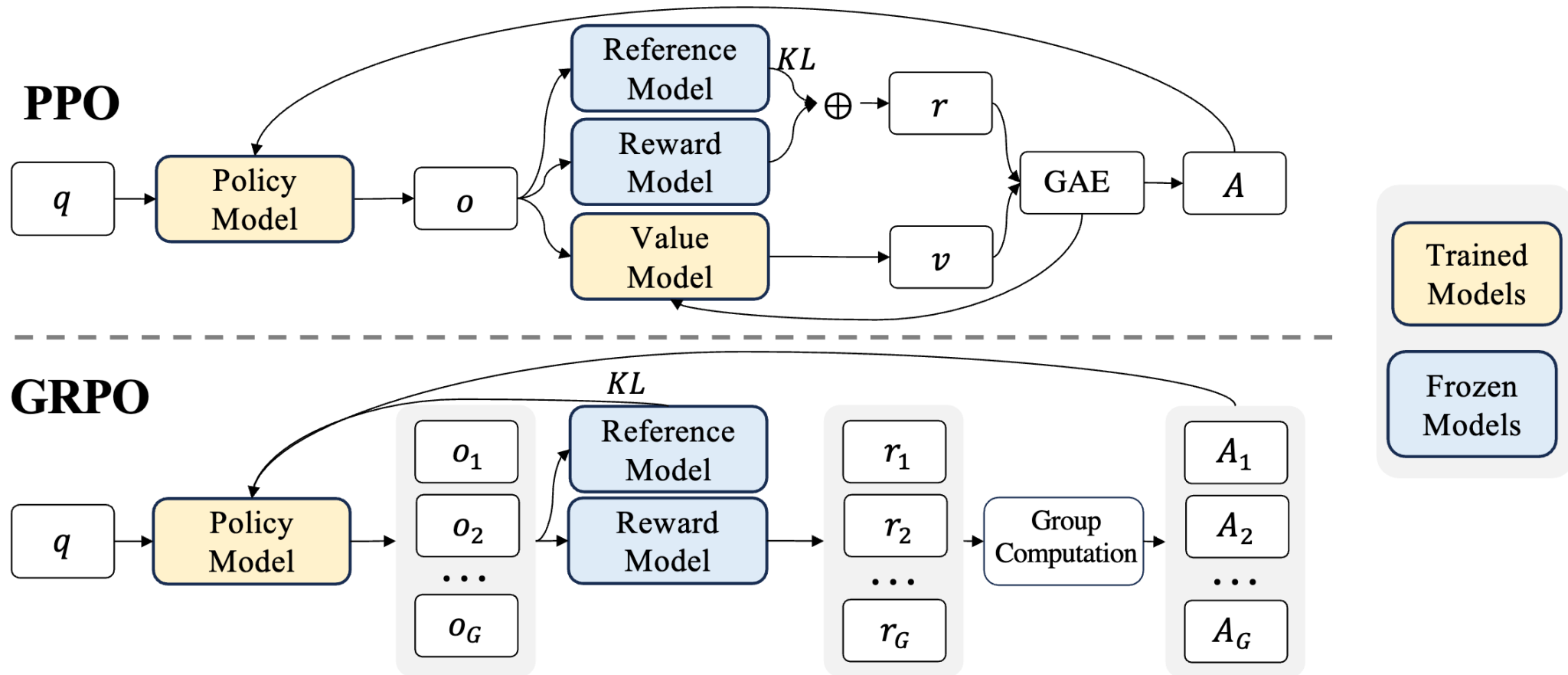
Method	Mistral-Base (7B)					Mistral-Instruct (7B)				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
SFT	8.4	6.2	1.3	4.8	6.3	17.1	14.7	12.6	6.2	7.5
RRHF [91]	11.6	10.2	5.8	5.4	6.7	25.3	24.8	18.1	6.5	7.6
SLiC-HF [96]	10.9	8.9	7.3	5.8	7.4	24.1	24.6	18.9	6.5	7.8
DPO [66]	15.1	12.5	10.4	5.9	7.3	26.8	24.9	16.3	6.3	7.6
IPO [6]	11.8	9.4	7.5	5.5	7.2	20.3	20.3	16.2	6.4	7.8
CPO [88]	9.8	8.9	6.9	5.4	6.8	23.8	28.8	22.6	6.3	7.5
KTO [29]	13.1	9.1	5.6	5.4	7.0	24.5	23.6	17.9	6.4	7.7
ORPO [42]	14.7	12.2	7.0	5.8	7.3	24.5	24.9	20.8	6.4	7.7
R-DPO [64]	17.4	12.8	8.0	5.9	7.4	27.3	24.5	16.1	6.2	7.5
SimPO	21.5	20.8	16.6	6.0	7.3	32.1	34.8	21.0	6.6	7.6

Method	Llama-3-Base (8B)					Llama-3-Instruct (8B)				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
SFT	6.2	4.6	3.3	5.2	6.6	26.0	25.3	22.3	6.9	8.1
RRHF [91]	12.1	10.1	6.3	5.8	7.0	31.3	28.4	26.5	6.7	7.9
SLiC-HF [96]	12.3	13.7	6.0	6.3	7.6	26.9	27.5	26.2	6.8	8.1
DPO [66]	18.2	15.5	15.9	6.5	7.7	40.3	37.9	32.6	7.0	8.0
IPO [6]	14.4	14.2	17.8	6.5	7.4	35.6	35.6	30.5	7.0	8.3
CPO [88]	10.8	8.1	5.8	6.0	7.4	28.9	32.2	28.8	7.0	8.0
KTO [29]	14.2	12.4	12.5	6.3	7.8	33.1	31.8	26.4	6.9	8.2
ORPO [42]	12.2	10.6	10.8	6.1	7.6	28.5	27.4	25.8	6.8	8.0
R-DPO [64]	17.6	14.4	17.2	6.6	7.5	41.1	37.8	33.1	7.0	8.0
SimPO	22.0	20.3	23.4	6.6	7.7	44.7	40.5	33.8	7.0	8.0



(c) Efficiency of DPO vs. SimPO.

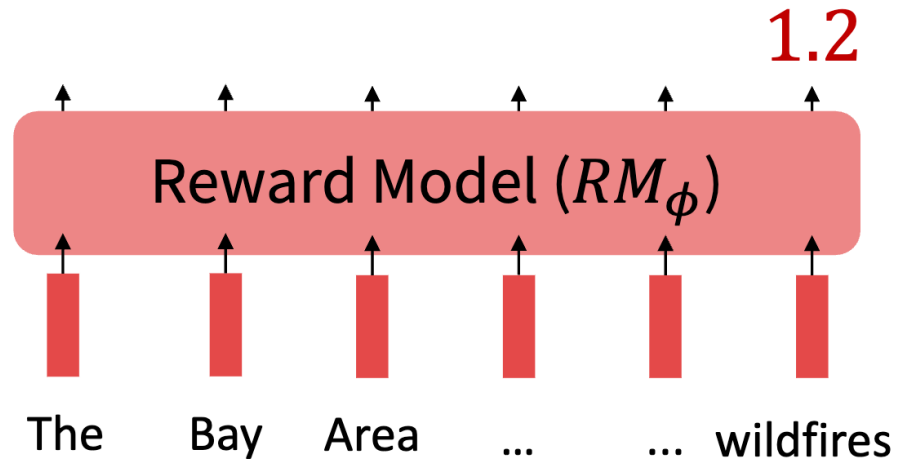
Group Relative Policy Optimization (GRPO)



Deepseek uses it!

Recap: Reward Model in PPO

- Train a reward model (RM) from an annotated dataset



$$\mathbb{E}_{\hat{s} \sim p_\theta(s)} [R(\hat{s})]$$

Group Relative Policy Optimization (GRPO)

- Consider **group relative** reward
 - Given x , sample multiple output y_1, y_2, \dots, y_G
 - Use reward model to get reward r_1, r_2, \dots, r_G

$$A_i = \frac{r_i - \text{mean}(r_1, r_2, \dots, r_G)}{\text{std}(r_1, r_2, \dots, r_G)}$$