



香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 4901B

Large Language Models

Instruction Tuning and Alignment

Junxian He

Oct 10, 2025

Review: Instruction Tuning vs Traditional Multi-task Fine-tuning

Machine learning wise, they are the same in terms of implementation

Review: Instruction Tuning vs Traditional Multi-task Fine-tuning

Machine learning wise, they are the same in terms of implementation

Traditional

Data is not that diverse,
typically 10s of tasks, each task
with >10K or even more
samples

Review: Instruction Tuning vs Traditional Multi-task Fine-tuning

Machine learning wise, they are the same in terms of implementation

Traditional

Data is not that diverse,
typically 10s of tasks, each task
with >10K or even more
samples

Instruction Tuning

Data is diverse, typically 1-3
examples per task, and
thousands of examples in total
can improve pretrained models
a lot

Review: Instruction Tuning vs Traditional Multi-task Fine-tuning

Machine learning wise, they are the same in terms of implementation

Traditional

Data is not that diverse, typically 10s of tasks, each task with >10K or even more samples

Instruction Tuning

Data is diverse, typically 1-3 examples per task, and thousands of examples in total can improve pretrained models a lot

What makes instruction tuning work with so few examples?

Review: Instruction Tuning vs Traditional Multi-task Fine-tuning

Supervised Finetuning (SFT)

Machine learning wise, they are the same in terms of implementation

Traditional

Data is not that diverse,
typically 10s of tasks, each task
with >10K or even more
samples

Instruction Tuning

Data is diverse, typically 1-3
examples per task, and
thousands of examples in total
can improve pretrained models
a lot

What makes instruction tuning work with so few examples?

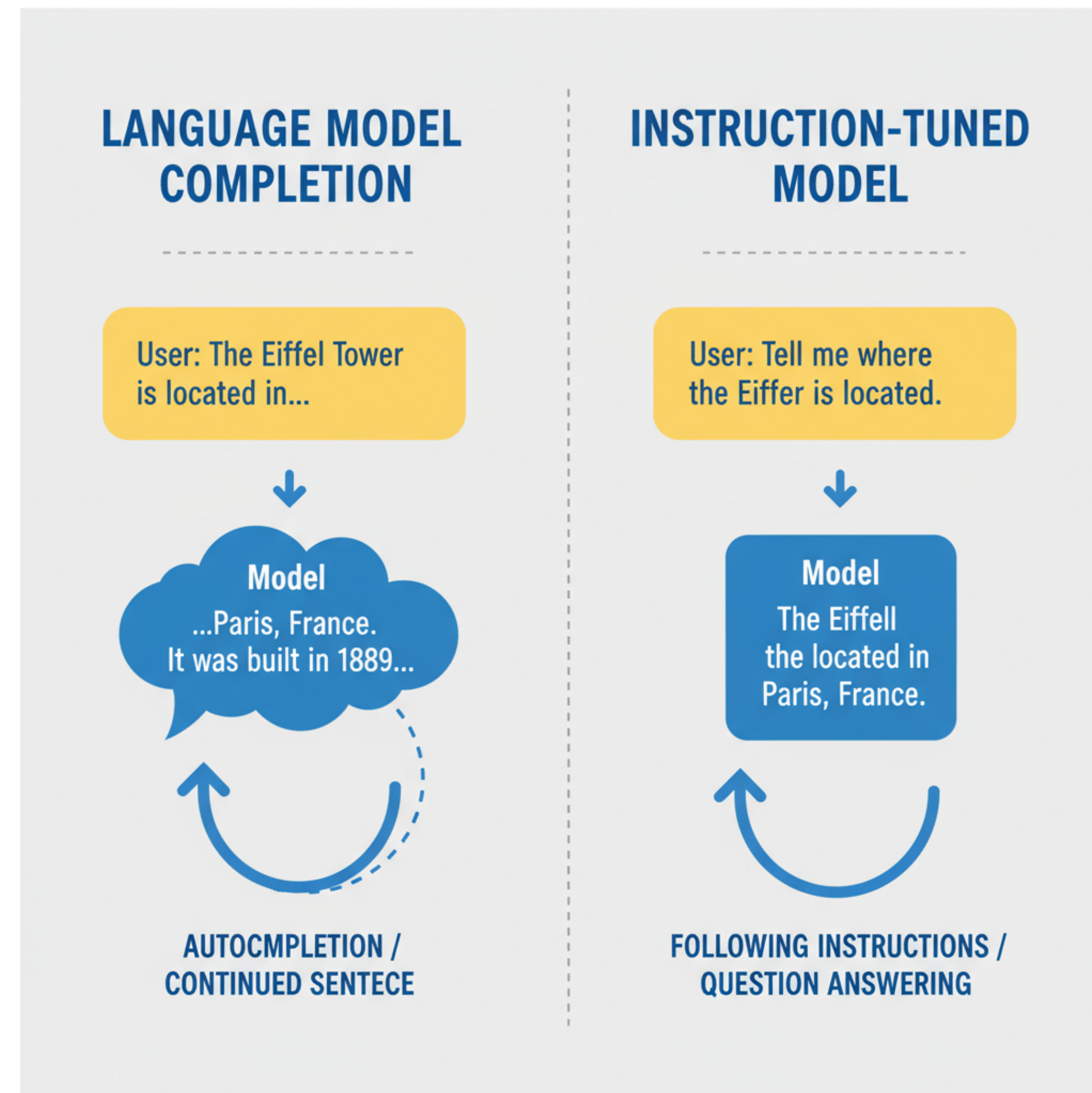
PRETRAINING

Review: (Human) Alignment

In a narrow definition, alignment means to adapt the language model to follow human instructions

Review: (Human) Alignment

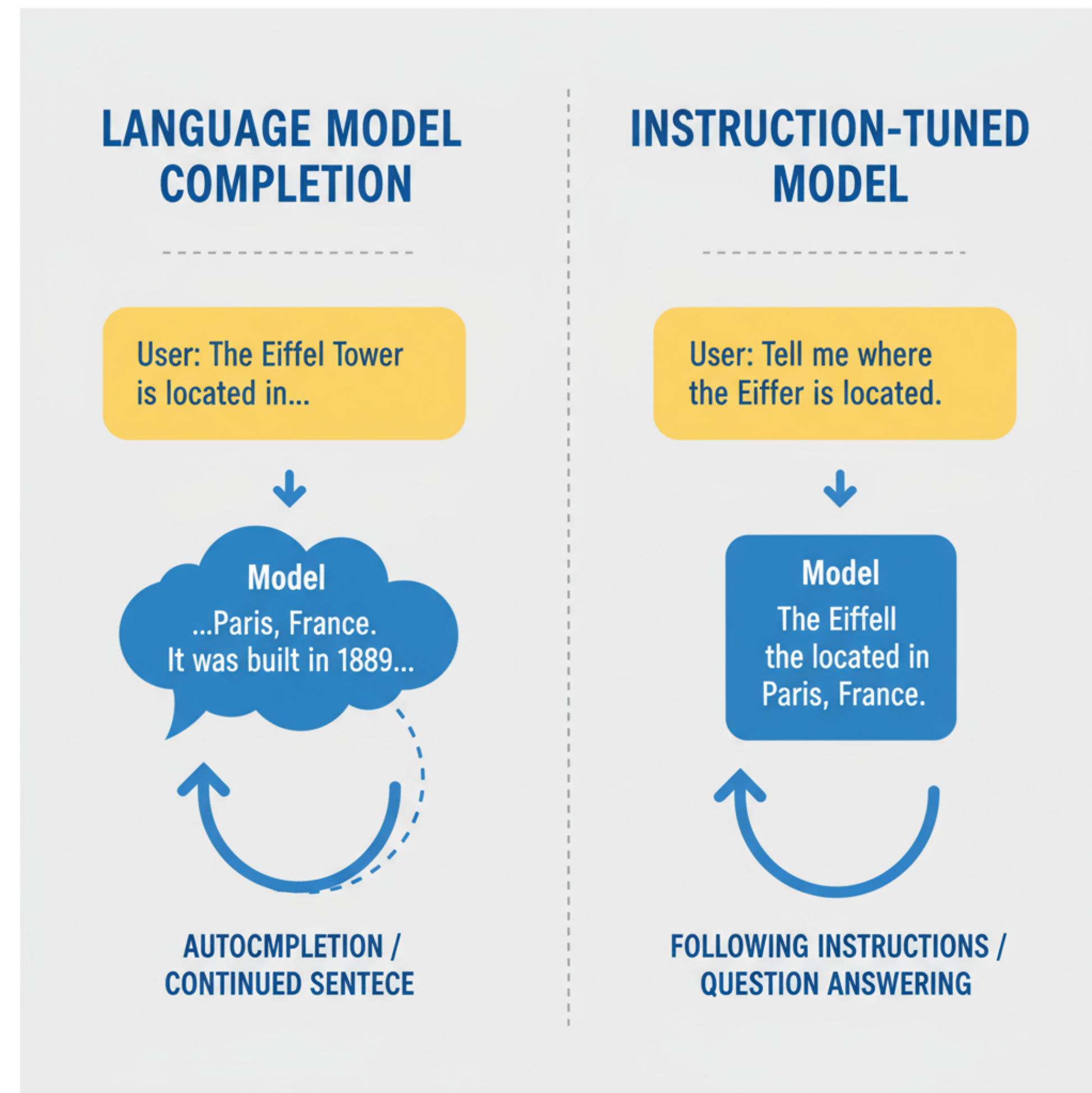
In a narrow definition, alignment means to adapt the language model to follow human instructions



Review: (Human) Alignment

In a narrow definition, alignment means to adapt the language model to follow human instructions

Sometimes, typical instruction tuning can be regarded as aligning the model



Review: (Human) Alignment

In a broad definition, alignment means to adapt the language model to align with human or society values, so that the models should not be toxic or biased (we'll cover safety aspects of LLMs later in this course)

Review: (Human) Alignment

In a broad definition, alignment means to adapt the language model to align with human or society values, so that the models should not be toxic or biased (we'll cover safety aspects of LLMs later in this course)



Instruction Tuning in Language Models



Large Language Models

What is the capital of France? The capital of France is Paris.

question

response

$[x, y]$

Instruction Tuning in Language Models

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

Instruction Tuning in Language Models

Left-shift one token as the output (because we are predicting the next token)

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models (masked attention)

What is the capital of France ? The capital of France is Paris .

Instruction Tuning in Language Models

Left-shift one token as the output (because we are predicting the next token)

is the capital of France ? The capital of France is Paris . 

Large Language Models

What is the capital of France ? The capital of France is Paris .

Note that the last token is a <stop> token so that the generation can automatically stop at test time

Instruction Tuning in Language Models

Left-shift one token as the output (because we are predicting the next token)

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

Note that the last token is a <stop> token so that the generation can automatically stop at test time

<stop> is not shown to users

Instruction Tuning in Language Models

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

Instruction Tuning in Language Models

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

$X_{1:m}$ = What is the capital of France ?

Instruction Tuning in Language Models

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

$X_{1:m} =$ What is the capital of France ?

$Y_{1:n} =$ The capital of France is Paris .<stop>

~~X~~ → Y

Instruction Tuning in Language Models

is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

What is the capital of France ? The capital of France is Paris .

$X_{1:m} =$ What is the capital of France ?

$Y_{1:n} =$ The capital of France is Paris .<stop>

$$L = - \sum_{i=1}^n \log p(Y_i | Y_{1:i-1}, X_{1:m})$$

all tokens on the left

& input

pretraining: loss loss loss

is the capital

-- The capital of --

L M

What is the capital -- ?

query/question

The capital of -- is --

answer

Instruction Tuning in Language Models

$X_{1:m} =$ What is the capital of France ?

$Y_{1:n} =$ The capital of France is Paris .<stop>

$$L = - \sum_{i=1}^n \log p(Y_i | Y_{1:i-1}, X_{1:m})$$

Difference from vanilla language modeling?

Instruction Tuning in Language Models

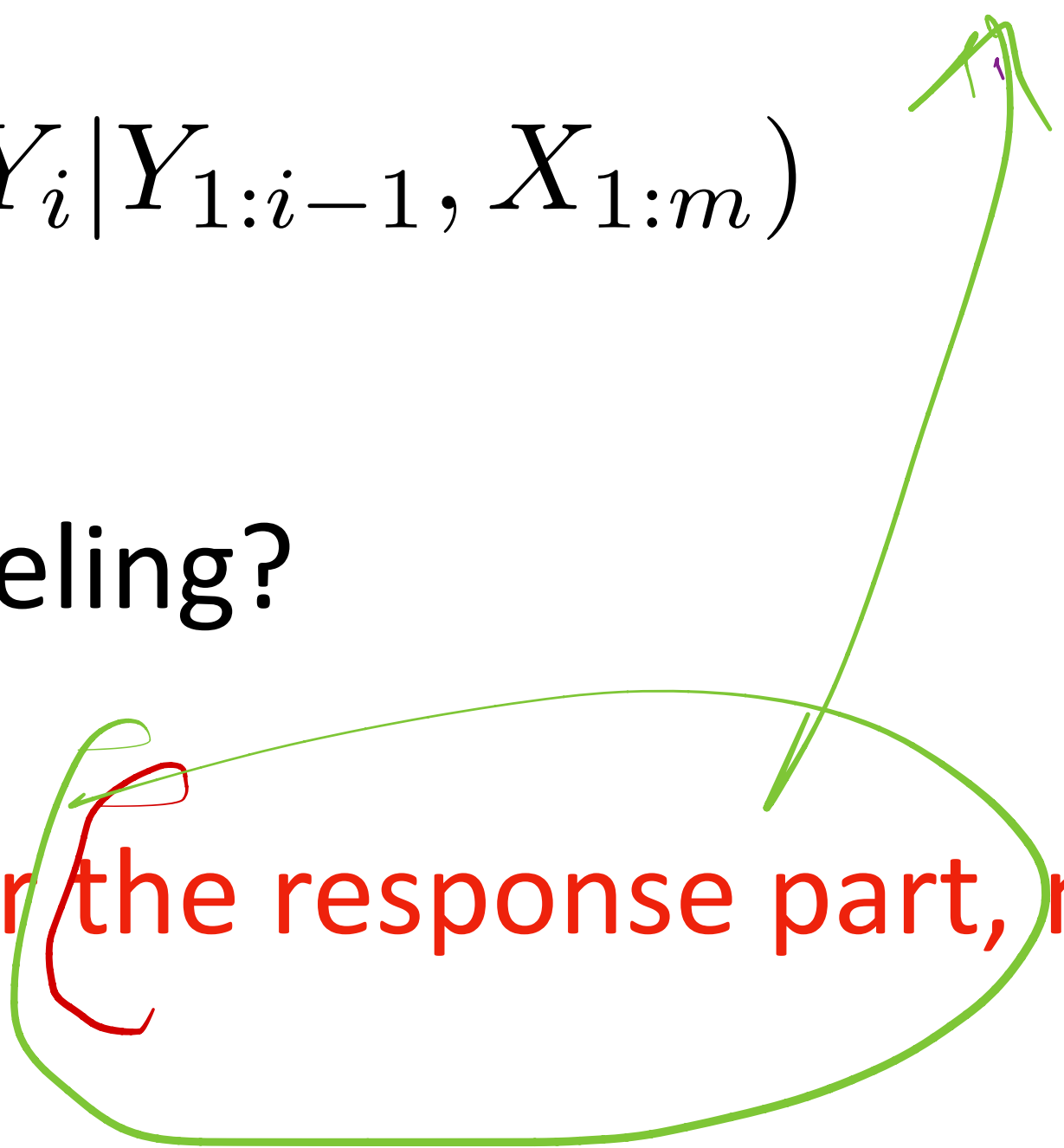
$X_{1:m} =$ What is the capital of France ?

$Y_{1:n} =$ The capital of France is Paris . <stop>

$$L = - \sum_{i=1}^n \log p(Y_i | Y_{1:i-1}, X_{1:m})$$

Difference from vanilla language modeling?

Only predicting the next token for the response part, not the input part



~~f f f f f~~

~~f f f f f~~

The capital of . . .

C M

What . . .

question

The capital of . . .

answer

(The capital of --- Y)

LM

x: What is the question ? ? ? ? ?

wrong: $P(Y_i | X_{1:m})$

correct: $P(Y_i | X_{1:m}, Y_{1:i-1})$

How to Implement?

How to Implement?

Input : **<start>** What is the capital of France ? **The capital**
of France is Paris .<stop>



How to Implement?

Input : <start> What is the capital of France ? The capital
of France is Paris . <stop>

Shift left by one position to get the output

How to Implement?

Input : <start> What is the capital of France ? The capital
of France is Paris .<stop>

Shift left by one position to get the output

Output : What is the capital of France ? The capital of
France is Paris .<stop>

How to Implement?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

In actual matrix computation, implementation is based on matrices, we compute the losses to every token on the output side

length = 10

(response) ↓

[1 2 x x x x x x x x]

hid hid
[] [] [] [] [] [] [] [] [] []

x x x x x x x x

length = 10

How to Implement?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

In actual matrix computation, implementation is based on matrices, we compute the losses to every token on the output side

Loss Masking

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

Loss Masking

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

loss list = [L(what), L(is), L(the),, L(Paris), L(.), L(<stop>)]

Mask list = [0, 0, 0,, 1, 1, 1]

Cross entropy loss = $\text{sum}(\text{loss_list} * \text{mask_list})$

0: question
1: response

pretraining:
 $\text{sum}(\text{loss list})$

Loss Masking

Sample code for loss masking

```
# -----  
logits = logits.view(-1, vocab_size)      # (batch*seq_len, vocab_size)  
labels = labels.view(-1)                 # (batch*seq_len,)  
mask = mask.view(-1)                     # (batch*seq_len,)  
  
# -----  
# 4 Compute cross-entropy loss per token  
# -----  
per_token_loss = F.cross_entropy(logits, labels, reduction='none')  
  
# -----  
# 5 Apply the mask  
# -----  
masked_loss = per_token_loss * mask  
  
# Mean only over response tokens  
loss = masked_loss.sum() / mask.sum()
```

loss list

Inference Time

Large Language Models

<start> What is the capital of France ?

This is your prompt



Inference Time

The

Large Language Models

<start> What is the capital of France ?

Inference Time

The capital

Large Language Models

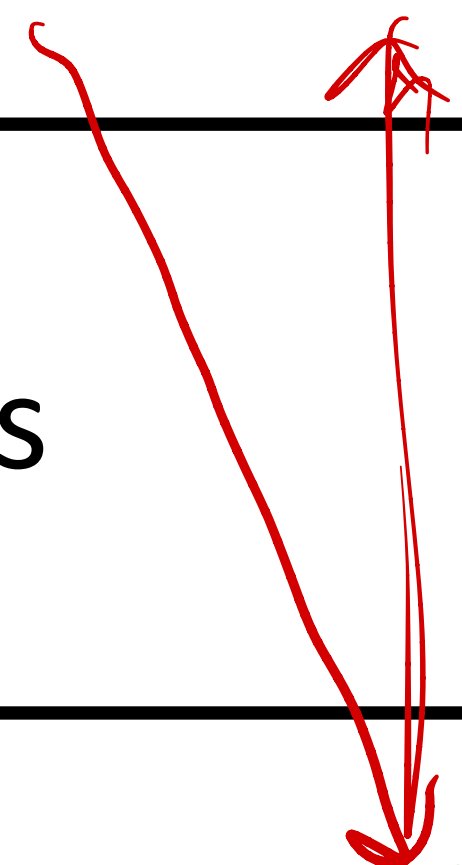
<start> What is the capital of France ? The

Inference Time

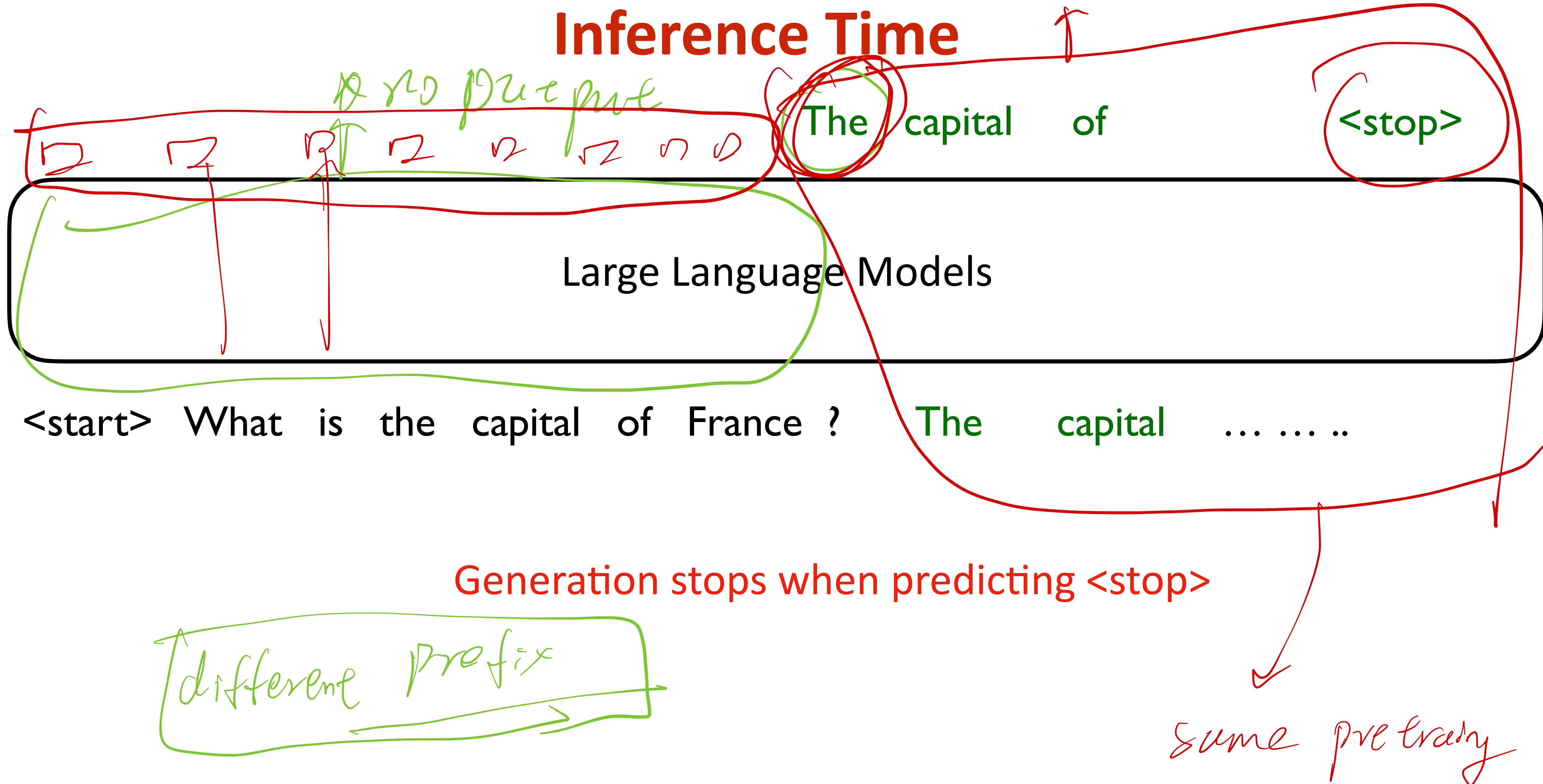
The capital of

Large Language Models

<start> What is the capital of France ? The capital



Inference Time



pretraining

STT:

A hand-drawn diagram of a horizontal capsule. A vertical line passes through the center of the capsule. The left side of the capsule is labeled 'präp' and the right side is labeled 'post'. The top of the capsule is labeled '12' and the bottom is labeled '11'.

why not:

$$P(Y_i | Y_{1:i-1})$$

$$P(\text{of} | \text{The, cap.})$$

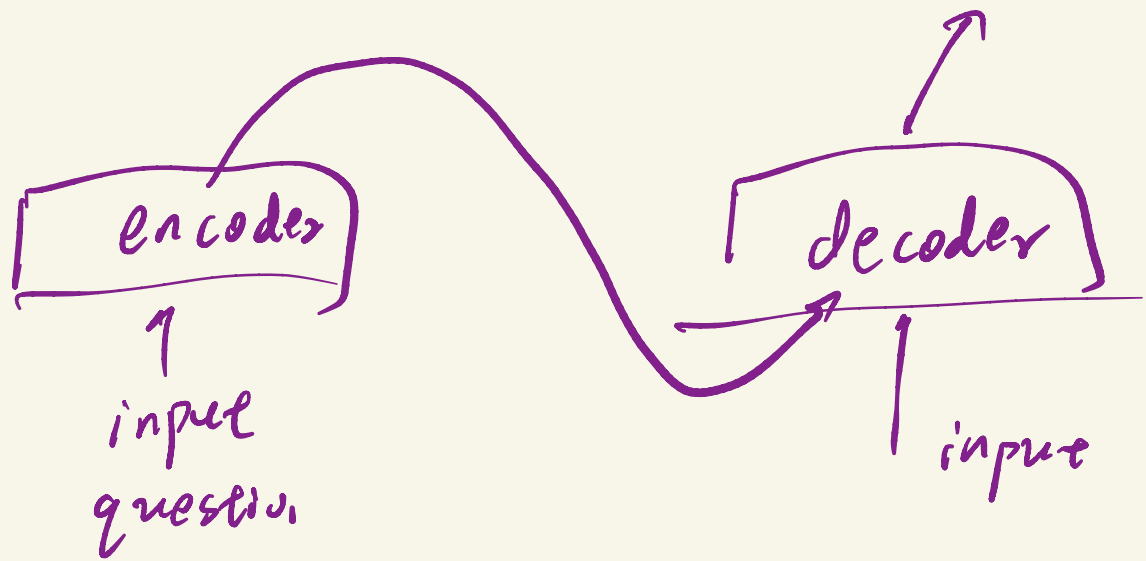
Y The capital of - - -

X



X What is the capital. -

✓ encoder The capital of .. decoder
What is the cap -



What if We don't do Loss Masking?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

What if We don't do Loss Masking?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

1. Can the model still answer instructions given prompt?
2. What else can the model do?



What if We don't do Loss Masking?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

1. Can the model still answer instructions given prompt?
2. What else can the model do?

The model can ask questions

What if We don't do Loss Masking?

What is the capital of France ? The capital of France is Paris . <stop>

Large Language Models

<start> What is the capital of France ? The capital of France is Paris .

1. Can the model still answer instructions given prompt?
2. What else can the model do?

The model can ask questions

It is ok to not mask for instruction tuning, which may or may not slightly hurt performance (no definitive conclusion)

Multi-Turn Instruction Tuning

<User>
<Assistant>

<step>

Conversational Data

<User> How are you today?

<Assistant> I'm doing well, thank you! How can I help you?

<User> Can you tell me a joke?

<Assistant> Sure! Why did the math book look sad? Because it had too many problems.

1st instruction

2nd instruction

Multi-Turn Instruction Tuning

Conversational Data

<User> How are you today?

<Assistant> I'm doing well, thank you! How can I help you?

<User> Can you tell me a joke?

<Assistant> Sure! Why did the math book look sad? Because it had too many problems.

When fed to language model, it becomes one sequence:

<User> How are you today? \n<Assistant> I'm doing well, thank you! How can I help you? \n\n<User> Can you tell me a joke? \n<Assistant> Sure! Why did the math book look sad?

Multi-Turn Instruction Tuning

<User> How are you today?\n<Assistant> I'm doing well, thank you! How can I help you? \n\n<User> Can you tell me a joke? \n<Assistant> Sure! Why did the math book look sad? <stop>

Large Language Models



<start> <User> How are you today?\n<Assistant> I'm doing well, thank you! How can I help you? \n\n<User> Can you tell me a joke? \n<Assistant> Sure! Why did the math book look sad?

Multi-Turn Instruction Tuning

<User> How are you today?\n<Assistant> I'm doing well, thank you! How can I help you? \n\n<User> Can you tell me a joke? \n<Assistant> Sure! Why did the math book look sad? <stop>

Large Language Models



<start> <User> How are you today?\n<Assistant> I'm doing well, thank you! How can I help you? \n\n<User> Can you tell me a joke? \n<Assistant> Sure! Why did the math book look sad?

No different from before, still shift one token left to obtain output

Multi-Turn Instruction Tuning

Large Language Models

model

<start> <User> How are you today? <user_stop> <Assistant> I'm doing well, thank you! How can I help you? <stop> <User> Can you tell me a joke?
<user_stop> <Assistant> Sure! Why did the math book look sad? <stop>

Stop token for each turn, so that each turn the model can automatically stop

why *template* *<bot>*

<User> <Assistant> tags are necessary

Inference time for ChatBot

Large Language Models



<start> <User> How are you today? <user_stop>

Your prompt

Inference time for ChatBot

Large Language Models

<start> <User> How are you today?<user_stop><Assistant>

Actual prompt

Inference time for ChatBot

Large Language Models



<start> <User> How are you today?<user_stop><Assistant>

Actual prompt

Typically, each model has certain “templates” to transform the sequence



Inference time for ChatBot

I'm doing well, thank you! How can I help you? <stop>

Large Language Models

<start> <User> How are you today? <user_stop> <Assistant>

Inference time for ChatBot

I'm doing well, thank you! How can I help you?<stop>

Large Language Models

<start> <User> How are you today?<user_stop><Assistant> I'm doing well, thank you! How can I help you?<stop>

Inference time for ChatBot

I'm doing well, thank you! How can I help you?<stop>

Large Language Models



<start> <User> How are you today?<user_stop><Assistant> I'm doing well, thank you! How can I help you?<stop> <User> Can you tell me a joke?
<user_stop><Assistant>



Inference time for ChatBot

I'm doing well, thank you! How can I help you? <stop>

Sure! Why did the math book look sad? <stop>

Large Language Models



<start> <User> How are you today? <user_stop> <Assistant> I'm doing well, thank you! How can I help you? <stop> <User> Can you tell me a joke?
<user_stop> <Assistant>

Inference time for ChatBot

I'm doing well, thank you! How can I help you? <stop>

Sure! Why did the math book look sad? <stop>

Large Language Models



<start> <User> How are you today? <user_stop> <Assistant> I'm doing well, thank you! How can I help you? <stop> <User> Can you tell me a joke?
<user_stop> <Assistant>

At inference time, we only ask the model to predict assistant parts

Inference time for ChatBot

How are you today?

I'm doing well, thank you! How can I help you? <stop>

Can you tell me a joke?

Sure! Why did the math book look sad? <stop>

Large Language Models

<start> <User> How are you today? <user_stop> <Assistant> I'm doing well, thank you! How can I help you? <stop> <User> Can you tell me a joke? <user_stop> <Assistant>

At inference time, we only ask the model to predict assistant parts

So typically, other parts are all masked when predicting the loss

mask = [0 0 0 0 0, ..., 1, 1, 1, 1, 1, ..., 0 0 0 0, 1 1 1 1]

Inference time for ChatBot

Large Language Models



<start> <User> How are you today?<user_stop><Assistant> I'm doing well, thank you! How can I help you?<stop> <User> Can you tell me a joke?
<user_stop><Assistant>

Or, we only mask <user> <assistant> tags, ~~not~~ learning all contents, then the chatbot can suggest questions each round

Chatbot can ask questions if not masking

Today

How are you today?
4:08 PM ✓

GPT-5-Chat

I'm doing great, thanks for asking! 😊
How about you? What's your day been like so far?
4:08 PM

Share ↻

Compare @o3-mini →

Compare @DeepSeek-R1-FW →

Compare @GPT-5 →

Speak @ElevenLabs-v2.5-Turbo →

It's been pretty good, just a regular day so far. →

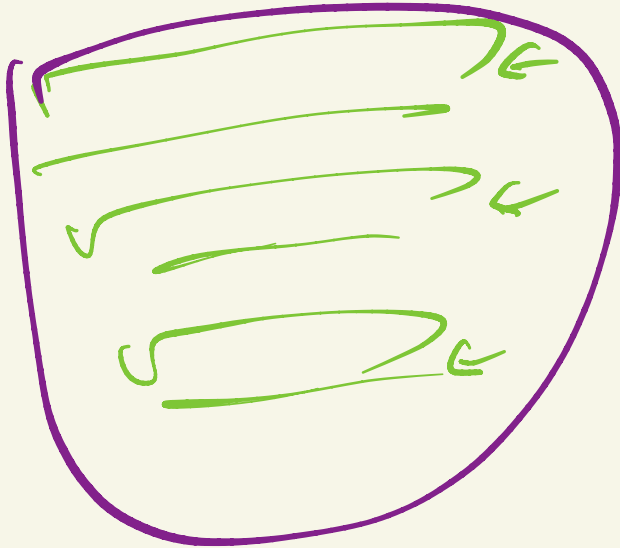
My day is going well, I'm just relaxing at the moment. →

It's been busy, but productive, how can I help you? →

no masking

$\langle \text{user} \rangle \times \dots \times \langle \text{assistant} \rangle \times \dots \times \langle \text{user} \rangle$

generate K different user inputs

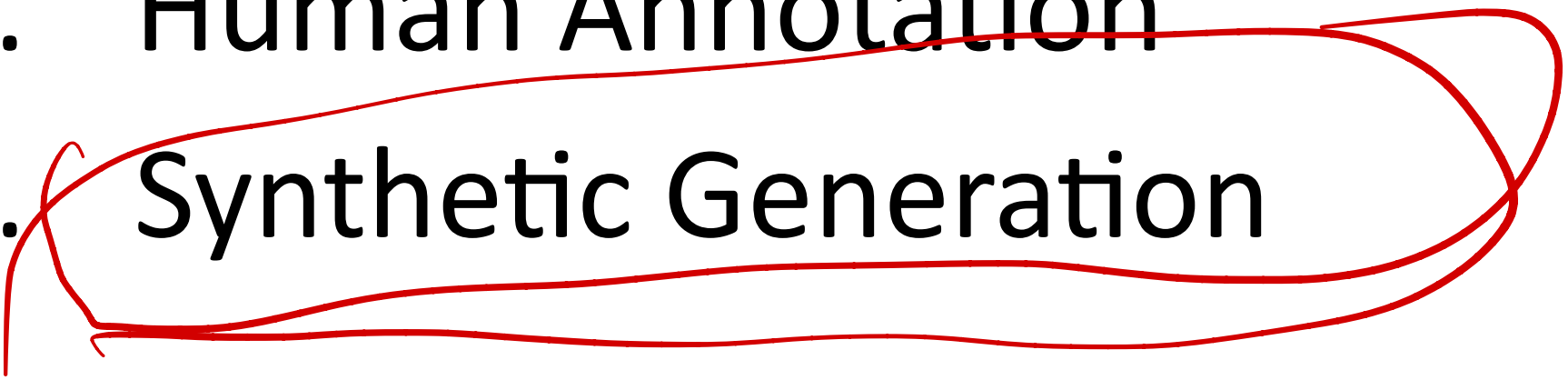


distribution

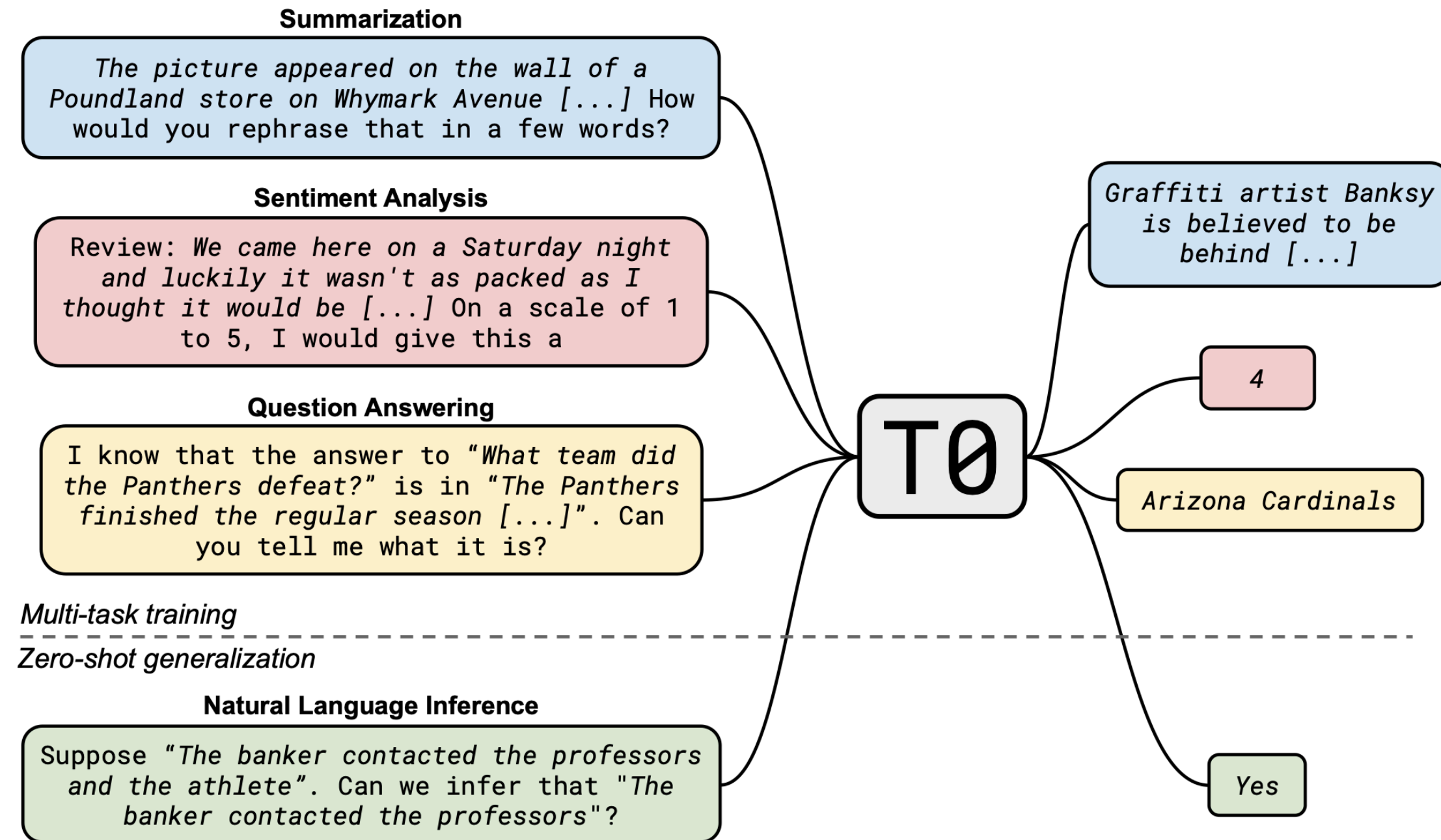
temperature sampling

↑
temperature

Where to get the Instruction Tuning Data

1. Source from existing NLP tasks
 2. Human Annotation
 3. Synthetic Generation
- 

Exiting NLP tasks



not diverse

<u>Natural language inference</u> (7 datasets) ANLI (R1-R3) RTE CB SNLI MNLI WNLI QNLI	<u>Commonsense</u> (4 datasets) CoPA HellaSwag PiQA StoryCloze	<u>Sentiment</u> (4 datasets) IMDB Sent140 SST-2 Yelp	<u>Paraphrase</u> (4 datasets) MRPC QQP PAWS STS-B	<u>Closed-book QA</u> (3 datasets) ARC (easy/chal.) NQ TQA	<u>Struct to text</u> (4 datasets) CommonGen DART E2ENLG WEBNLG	<u>Translation</u> (8 datasets) ParaCrawl EN/DE ParaCrawl EN/ES ParaCrawl EN/FR WMT-16 EN/CS WMT-16 EN/DE WMT-16 EN/FI WMT-16 EN/RO WMT-16 EN/RU WMT-16 EN/TR
<u>Reading comp.</u> (5 datasets) BoolQ OBQA DROP SQuAD MultiRC	<u>Read. comp. w/ commonsense</u> (2 datasets) CosmosQA ReCoRD	<u>Coreference</u> (3 datasets) DPR Winogrande WSC273	<u>Misc.</u> (7 datasets) CoQA TREC QuAC CoLA WIC Math Fix Punctuation (NLG)	<u>Summarization</u> (11 datasets) AESLC Multi-News SamSum AG News Newsroom Wiki Lingua EN CNN-DM Opin-Abs: iDebate XSum Gigaword Opin-Abs: Movie		

Human Annotation

chat GPT

SFT
instruction
every

Step 1
Collect demonstration data
and train a supervised policy.

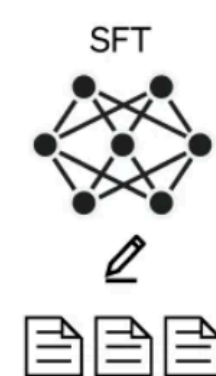
A prompt is sample from
our prompt dataset.

Explain reinforcement
learning to a 6 year old.

A labeler demonstrates
the desired output
behavior.

We give treats and
punishments to teach...

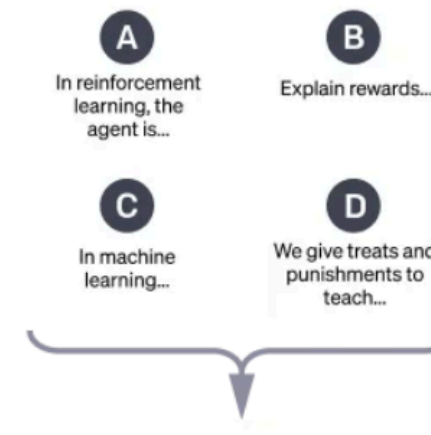
This data is used to
fine-tune GPT-3.5 with
supervised learning.



Step 2
Collect comparison data and
train a reward model.

A prompt and several
model outputs are
sampled.

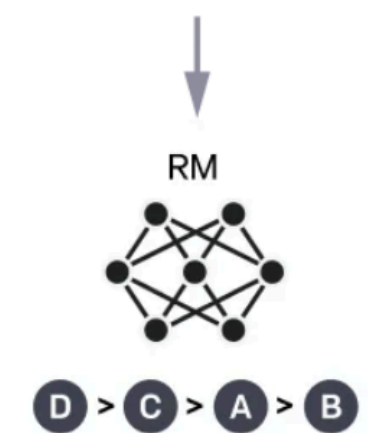
Explain reinforcement
learning to a 6 year old.



A labeler ranks the
outputs from best
to worst.

D > C > A > B

This data is used to
train our reward model.



Step 3
Optimize a policy against the
reward model using the PPO
reinforcement learning algorithm.

A new prompt is
sampled from
the dataset.

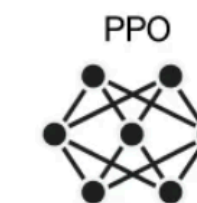
The PPO model is
initialized from the
supervised policy.

The policy generates
an output.

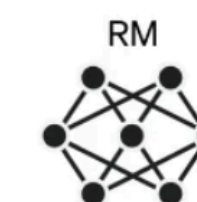
The reward model
calculates a reward
for the output.

The reward is used
to update the policy
using PPO.

Write a story
about otters.



Once upon a time...



r_k

prompt dataset?

from users

Human Annotation

Step 1

Collect demonstration data and train a supervised policy.

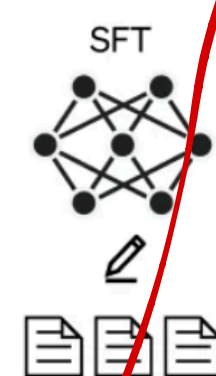
A prompt is sample from our prompt dataset.

Explain reinforcement learning to a 6 year old.

A labeler demonstrates the desired output behavior.

We give treats and punishments to teach...

This data is used to fine-tune GPT-3.5 with supervised learning.

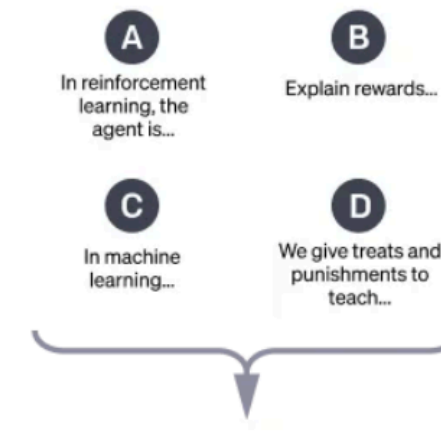


Step 2

Collect comparison data and train a reward model.

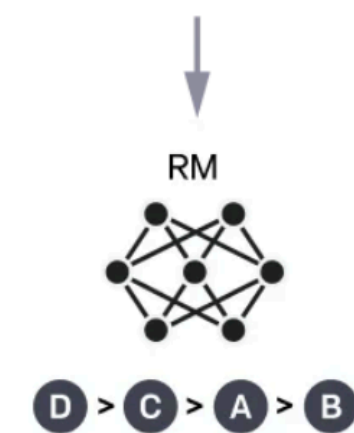
A prompt and several model outputs are sampled.

Explain reinforcement learning to a 6 year old.



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 the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

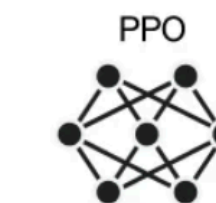
Write a story about otters.

The PPO model is initialized from the supervised policy.

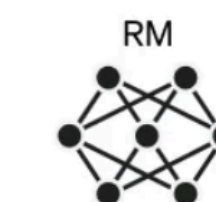
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Once upon a time...



r_k

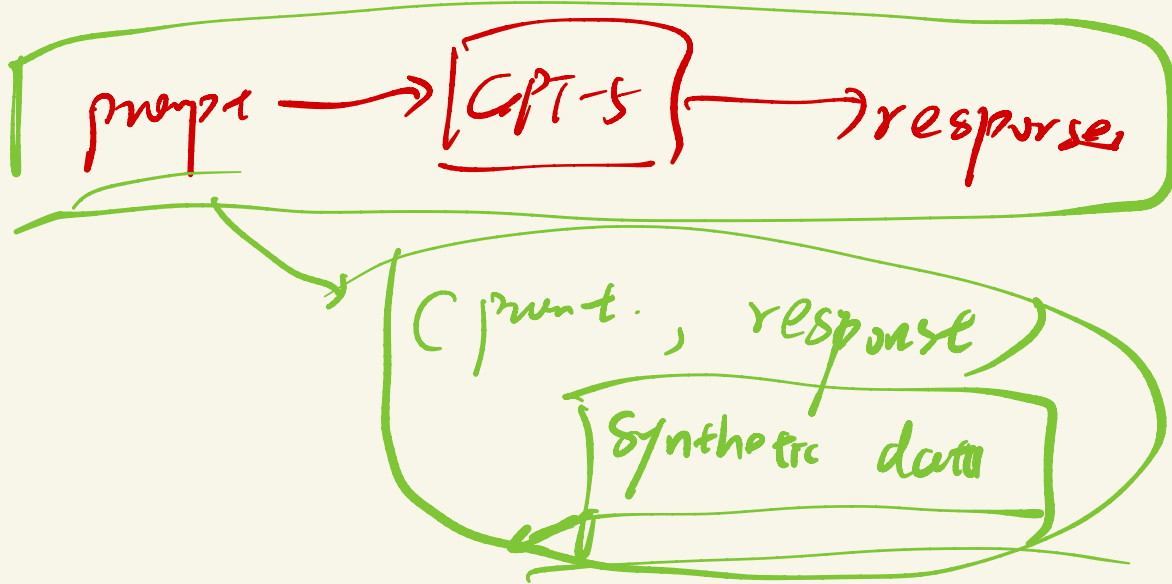
most reliable way

ChatGPT uses human annotation in Step 1

Synthetic data:

prompt data

where to get responses?



synthetic data is unavoidable

web → pretraining data



most of them



human data ?

synthetic data

will be synthetic data some day

Thank You!