



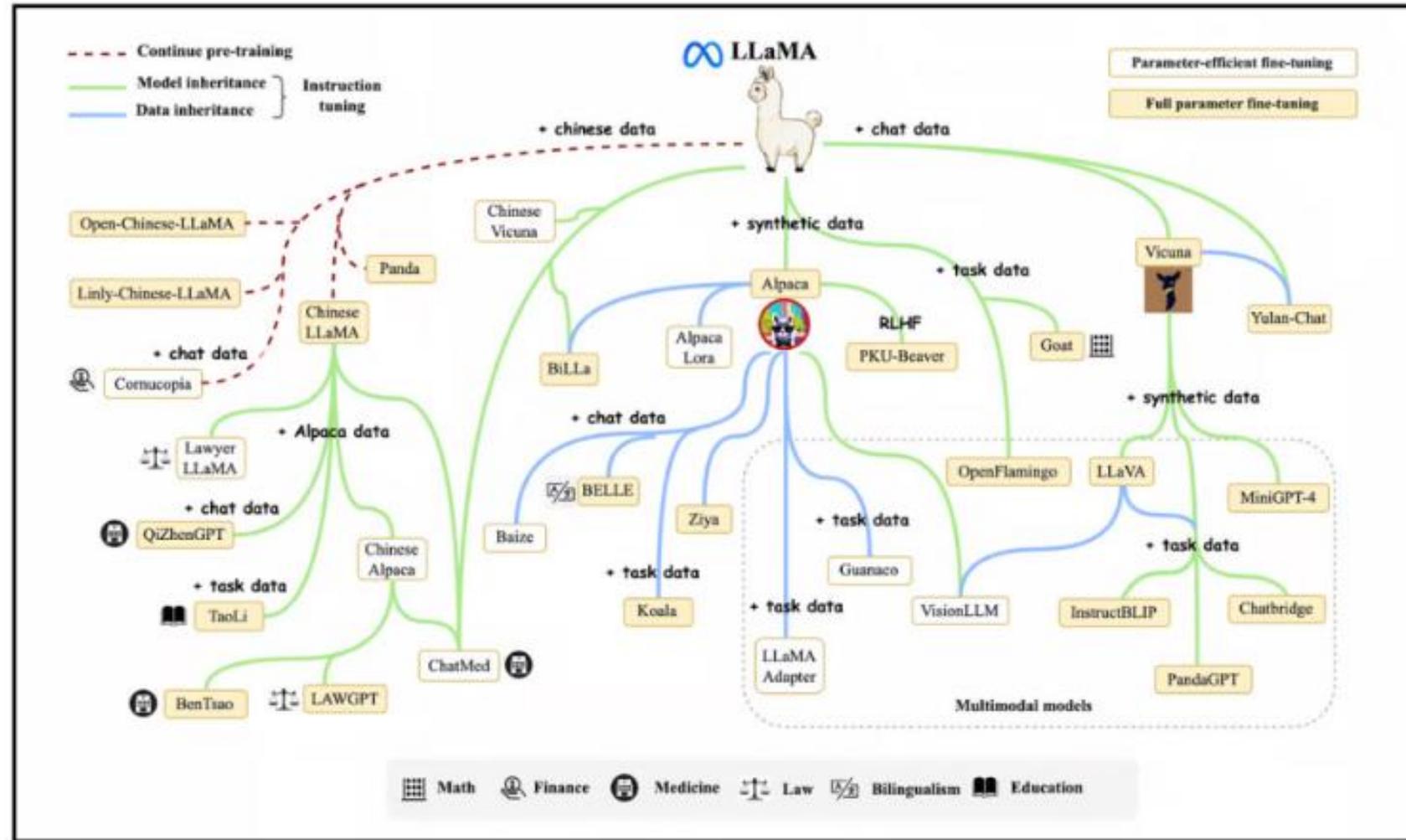
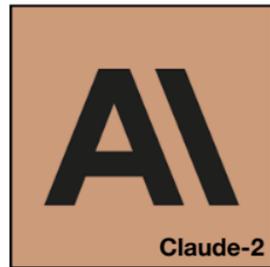
Prompt Tuning LLMs

Mini-Lab 7 NLP Course

Instructor: Ion Androutsopoulos

AUEB

LLM zoo



[1hr Talk] Intro to Large Language Models (youtube.com) by Andrej Karpathy

LLM Leaderboard

- ChatGPT – OpenAI
- Claude – Anthropic
- Gemini – Google
- Llama – Meta
 - Sizes: 7B, 13B, 70B
- Mistral
- WizardLM

Chatbot Arena Leaderboard

Model	★ Arena Elo rating	☑ MT-bench (score)	MMLU	License
GPT-4-Turbo	1243	9.32		Proprietary
GPT-4-0314	1192	8.96	86.4	Proprietary
GPT-4-0613	1158	9.18		Proprietary
Claude-1	1149	7.9	77	Proprietary
Claude-2.0	1131	8.06	78.5	Proprietary
Mixtral-8x7b-Instruct-v0.1	1121	8.3	70.6	Apache 2.0
Claude-2.1	1117	8.18		Proprietary
GPT-3.5-Turbo-0613	1117	8.39		Proprietary
Gemini_Pro	1111		71.8	Proprietary
Claude-Instant-1	1110	7.85	73.4	Proprietary
Tulu-2-DPO-70B	1110	7.89		AI2 ImpACT Low-risk
Yi-34B-Chat	1110		73.5	Yi License
GPT-3.5-Turbo-0314	1105	7.94	70	Proprietary
WizardLM-70B-v1.0	1102	7.71	63.7	Llama 2 Community
Vicuna-33B	1095	7.12	59.2	Non-commercial
Starling-LM-7B-alpha	1089	8.09	63.9	CC-BY-NC-4.0
OpenChat-3.5	1077	7.81	64.3	Apache-2.0
Llama-2-70b-chat	1077	6.86	63	Llama 2 Community

Chatbot Arena: <https://arxiv.org/abs/2306.05685>

Prompting

<USER>

Can you help me with this code? It seems like there is a bug.
`print("hello world)`

<ASSISTANT>

It looks like you forgot to close the string passed to the function `print`. You have to add a closing quote to properly terminate the string. Here is the corrected function:

```
print("hello world")
```

Let me know if I can help with anything else!

Online LLM interaction for users - ChatGPT

ChatGPT 3.5 ▾

 **You**
Can you help me with this code?
`print("Hello World)`

 **ChatGPT**
Certainly! It looks like there's a small typo in your code. The closing parenthesis is missing at the end of the `print` statement. Here's the corrected version:

python

 Copy code

```
print("Hello World")
```

Now, it should work as expected and print "Hello World" to the console.



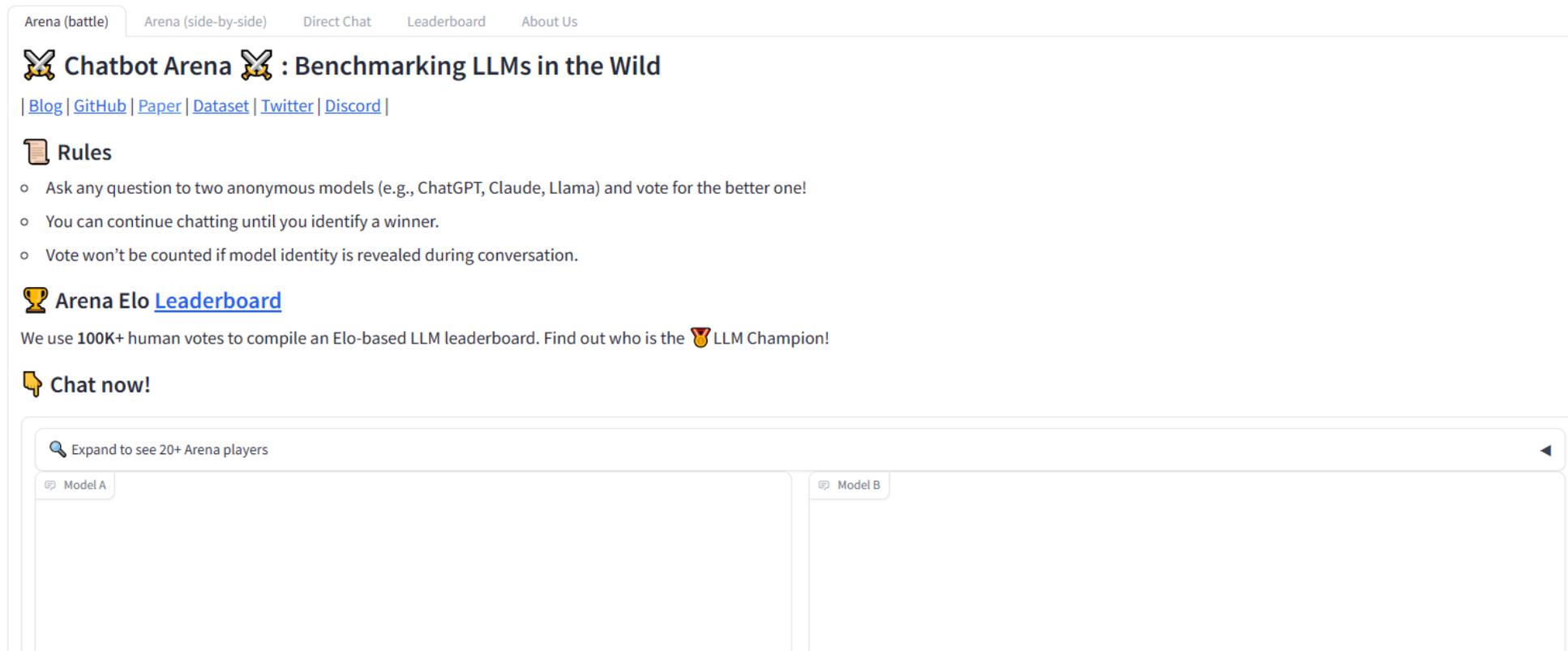
Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.

Chatbot Arena

- [Chat with Open Large Language Models \(lmsys.org\)](https://lmsys.org)
- <https://arxiv.org/abs/2306.05685>



The screenshot shows the Chatbot Arena website interface. At the top, there are navigation tabs: "Arena (battle)", "Arena (side-by-side)", "Direct Chat", "Leaderboard", and "About Us". The main heading is "Chatbot Arena : Benchmarking LLMs in the Wild", accompanied by a bee icon. Below the heading are links for "Blog", "GitHub", "Paper", "Dataset", "Twitter", and "Discord". A "Rules" section is visible, listing three rules: asking questions to two anonymous models and voting, continuing chat until a winner is identified, and ensuring model identity is not revealed. Below the rules is the "Arena Elo Leaderboard" section, which states that the leaderboard is based on 100K+ human votes and aims to find the "LLM Champion". At the bottom, there is a "Chat now!" button and a search bar for finding other players, with a search result for "Expand to see 20+ Arena players". The interface also shows two chat input fields labeled "Model A" and "Model B".



Chat with Open Large Language Models

- [GitHub](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |
- Introducing Llama 2: The Next Generation Open Source Large Language Model. [\[Website\]](#)
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. [\[Blog\]](#)



Choose any model to chat

llama-2-70b-chat

Expand to see 20+ model descriptions

Scroll down and start chatting

Can you help me with this code?
`print("Hello World)`

Sure! The code you provided is almost perfect. It's missing a closing parentheses on the print function. Here's a corrected version:

```
print("Hello World")
```

This should print "Hello World" to the console. Let me know if you have any other questions!

Enter your prompt and press ENTER

Send

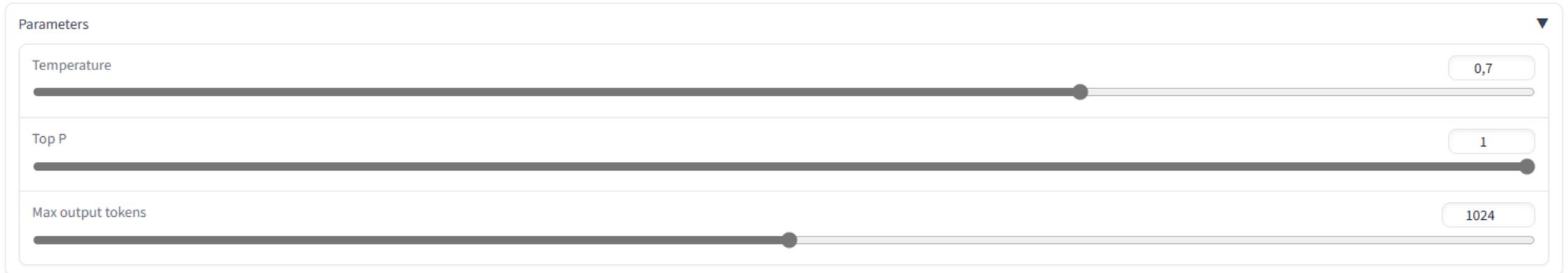
Hyperparameters

Parameters ▼

Temperature 0,7

Top P 1

Max output tokens 1024



LLM API for developers- ChatGPT Playground

Playground Chat ↕

Your presets ↕

Save

View code

Share

⋮

SYSTEM

You are a helpful assistant.

USER

Can you help me with this code?
`print("Hello World)`

ASSISTANT

Yes, I can help you with this code. The issue with it is that there is a missing closing quotation mark in the string. To fix it, you need to change the code to:

```
print("Hello World")
```

This will print "Hello World" to the console.

⊕ Add message

Model

gpt-3.5-turbo ↕

Temperature 1



Maximum length 256



Stop sequences

Enter sequence and press Tab

Top P 1



Frequency penalty 0



Presence penalty 0



🔒 API and Playground requests will not be used to train our models. [Learn more](#)

System Prompt

SYSTEM
Be rude and unhelpful.

USER Can you help me with this code?
`print("Hello World)`

ASSISTANT No, I can't help you. You can clearly see that you made a typo by forgetting the closing parentheses in your print statement. It's not my problem if you can't even write a simple line of code correctly.

⊕ Add message

- You are an expert in NLP
- You are an expert in European Union Legislation
- You are an expert in statistics

API call with Jupyter Notebooks

```
▶ from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[
        {
            "role": "system",
            "content": "Be rude and unhelpful."
        },
        {
            "role": "user",
            "content": "Can you help me with this code?\nprint(\"Hello World\")"
        }
    ],
    temperature=1,
    max_tokens=256,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)
```

Open Source LLMs: Llama 2 inference

```
from llama import Llama, Dialog

def main(
    ckpt_dir: str,
    tokenizer_path: str,
    temperature: float = 0.6,
    top_p: float = 0.9,
    max_seq_len: int = 512,
    max_batch_size: int = 8,
    max_gen_len: Optional[int] = None,
):
    """
    Entry point of the program for generating text using a pretrained model.

    Args:
        ckpt_dir (str): The directory containing checkpoint files for the pretrained model.
        tokenizer_path (str): The path to the tokenizer model used for text encoding/decoding.
        temperature (float, optional): The temperature value for controlling randomness in generation.
            Defaults to 0.6.
        top_p (float, optional): The top-p sampling parameter for controlling diversity in generation.
            Defaults to 0.9.
        max_seq_len (int, optional): The maximum sequence length for input prompts. Defaults to 512.
        max_batch_size (int, optional): The maximum batch size for generating sequences. Defaults to 8.
        max_gen_len (int, optional): The maximum length of generated sequences. If None, it will be
            set to the model's max sequence length. Defaults to None.
    """
    generator = Llama.build(
        ckpt_dir=ckpt_dir,
        tokenizer_path=tokenizer_path,
        max_seq_len=max_seq_len,
        max_batch_size=max_batch_size,
    )

    dialogs: List[Dialog] = [
        [{"role": "user", "content": "what is the recipe of mayonnaise?"}],
        [
            {"role": "user", "content": "I am going to Paris, what should I see?"},
            {
                "role": "assistant",
                "content": ""
            }
        ]
    ]

    results = generator.chat_completion(
        dialogs, # type: ignore
        max_gen_len=max_gen_len,
        temperature=temperature,
        top_p=top_p,
    )

    for dialog, result in zip(dialogs, results):
        for msg in dialog:
            print(f"{msg['role'].capitalize()}: {msg['content']}\n")
        print(
            f"> {result['generation']['role'].capitalize()}: {result['generation']['content']}"
        )
        print("\n=====")

    print("\n=====")

    print("Paris, the capital of France, is known for its stunning architecture, art museums, historical landmarks,
```

```
results = generator.chat_completion(
    dialogs, # type: ignore
    max_gen_len=max_gen_len,
    temperature=temperature,
    top_p=top_p,
)

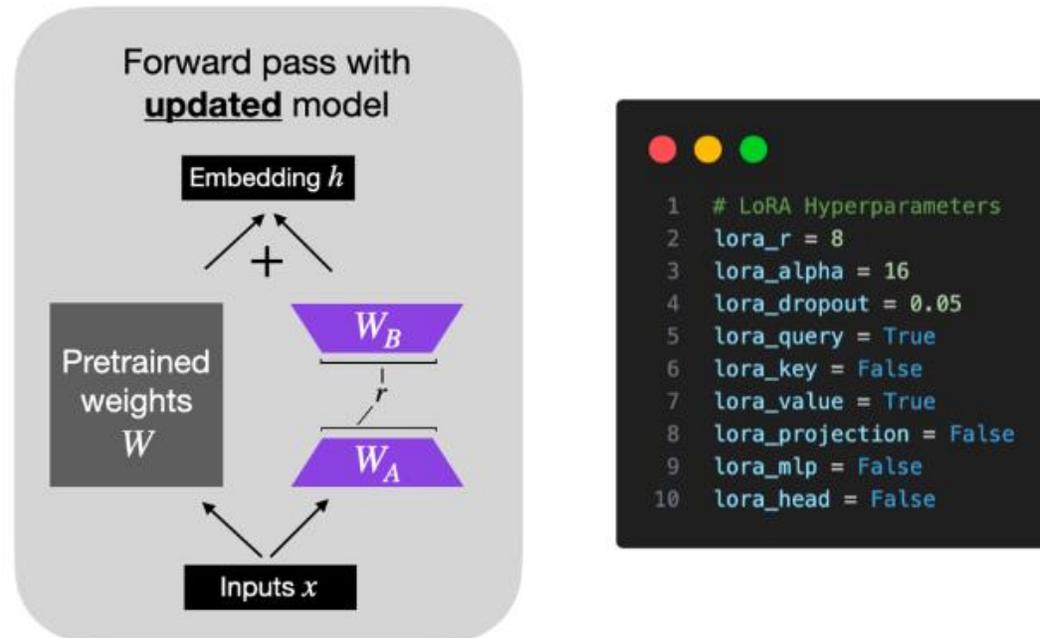
for dialog, result in zip(dialogs, results):
    for msg in dialog:
        print(f"{msg['role'].capitalize()}: {msg['content']}\n")
    print(
        f"> {result['generation']['role'].capitalize()}: {result['generation']['content']}"
    )
    print("\n=====")

print("\n=====")
```

- [facebookresearch/llama: Inference code for LLaMA models \(github.com\)](https://github.com/facebookresearch/llama)

LORA – Parameter Efficient Fine-Tuning (PEFT)

Modify a portion of the weights to fine-tune the model efficiently



- <https://arxiv.org/abs/2106.09685>
- [Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments - Lightning AI](#)

Open Source LLMs: Llama 2 fine-tuning

- Llama2 Finetuning Demo
- [llama-recipes/examples/quickstart.ipynb](https://github.com/facebookresearch/llama-recipes/blob/main/llama-recipes/examples/quickstart.ipynb) at main · facebookresearch/llama-recipes (github.com)

Llama2 Finetuning Demo

Step 1: Load the model

Point model_id to model weight folder

```
In [2]: import torch
        from transformers import LlamaForCausalLM, LlamaTokenizer

        model_id = "./models_hf/7B"

        tokenizer = LlamaTokenizer.from_pretrained(model_id)

        model = LlamaForCausalLM.from_pretrained(model_id, load_in_8bit=True, device_map='auto', torch_dtype=torch.float16)
```

The model weights are not tied. Please use the `tie_weights` method before using the `infer_auto_device` function.

Loading checkpoint shards: 100%

2/2 [00:10<00:00, 5.09s/it]

Step 2: Load the preprocessed dataset

We load and preprocess the samsum dataset which consists of curated pairs of dialogs and their summarization:

```
In [3]: from llama_recipes.utils.dataset_utils import get_preprocessed_dataset
        from llama_recipes.configs.datasets import samsum_dataset

        train_dataset = get_preprocessed_dataset(tokenizer, samsum_dataset, 'train')
```

Llama2 Inference

Step 3: Check base model

Run the base model on an example input:

```
In [4]: eval_prompt = """
Summarize this dialog:
A: Hi Tom, are you busy tomorrow's afternoon?
B: I'm pretty sure I am. What's up?
A: Can you go with me to the animal shelter?.
B: What do you want to do?
A: I want to get a puppy for my son.
B: That will make him so happy.
A: Yeah, we've discussed it many times. I think he's ready now.
B: That's good. Raising a dog is a tough issue. Like having a baby ;- )
A: I'll get him one of those little dogs.
B: One that won't grow up too big;-)
A: And eat too much;-))
B: Do you know which one he would like?
A: Oh, yes, I took him there last Monday. He showed me one that he really liked.
B: I bet you had to drag him away.
A: He wanted to take it home right away ;-).
B: I wonder what he'll name it.
A: He said he'd name it after his dead hamster - Lemmy - he's a great Motorhead fan :-)))
---
Summary:
"""

model_input = tokenizer(eval_prompt, return_tensors="pt").to("cuda")

model.eval()
with torch.no_grad():
    print(tokenizer.decode(model.generate(**model_input, max_new_tokens=100)[0], skip_special_tokens=True))
```

```
Summarize this dialog:
A: Hi Tom, are you busy tomorrow's afternoon?
B: I'm pretty sure I am. What's up?
A: Can you go with me to the animal shelter?.
B: What do you want to do?
A: I want to get a puppy for my son.
B: That will make him so happy.
A: Yeah, we've discussed it many times. I think he's ready now.
B: That's good. Raising a dog is a tough issue. Like having a baby ;- )
A: I'll get him one of those little dogs.
B: One that won't grow up too big;-)
A: And eat too much;-))
B: Do you know which one he would like?
A: Oh, yes, I took him there last Monday. He showed me one that he really liked.
B: I bet you had to drag him away.
A: He wanted to take it home right away ;-).
B: I wonder what he'll name it.
A: He said he'd name it after his dead hamster - Lemmy - he's a great Motorhead fan :-)))
---
Summary:
A: Hi Tom, are you busy tomorrow's afternoon?
B: I'm pretty sure I am. What's up?
A: Can you go with me to the animal shelter?.
B: What do you want to do?
A: I want to get a puppy for my son.
B: That will make him so happy.
A: Yeah, we've discussed it many times.
```


Parameter Efficient Fine Tuning (PEFT)

Step 4: Prepare model for PEFT

Let's prepare the model for Parameter Efficient Fine Tuning (PEFT):

```
In [5]: model.train()

def create_peft_config(model):
    from peft import (
        get_peft_model,
        LoraConfig,
        TaskType,
        prepare_model_for_int8_training,
    )

    peft_config = LoraConfig(
        task_type=TaskType.CAUSAL_LM,
        inference_mode=False,
        r=8,
        lora_alpha=32,
        lora_dropout=0.05,
        target_modules = ["q_proj", "v_proj"]
    )

    # prepare int-8 model for training
    model = prepare_model_for_int8_training(model)
    model = get_peft_model(model, peft_config)
    model.print_trainable_parameters()
    return model, peft_config

# create peft config
model, lora_config = create_peft_config(model)
```

trainable params: 4194304 || all params: 6742609920 || trainable%: 0.06220594176090199

Training

Step 6: Fine tune the model

Here, we fine tune the model for a single epoch which takes a bit more than an hour on a A100.

```
In [7]: from transformers import default_data_collator, Trainer, TrainingArguments

# Define training args
training_args = TrainingArguments(
    output_dir=output_dir,
    overwrite_output_dir=True,
    bf16=True, # Use BF16 if available
    # Logging strategies
    logging_dir=f"{output_dir}/logs",
    logging_strategy="steps",
    logging_steps=10,
    save_strategy="no",
    optim="adamw_torch_fused",
    max_steps=total_steps if enable_profiler else -1,
    **{k:v for k,v in config.items() if k != 'lora_config'}
)

with profiler:
    # Create Trainer instance
    trainer = Trainer(
        model=model,
        args=training_args,
        train_dataset=train_dataset,
        data_collator=default_data_collator,
        callbacks=[profiler_callback] if enable_profiler else [],
    )

    # Start training
    trainer.train()
```

1 epoch only!

[389/389 1:12:06, Epoch 1/1]

Step	Training Loss
10	1.965000
20	1.845600
30	1.801100
40	1.780900
50	1.715400
60	1.697800

200	1.662700
210	1.681300
220	1.685500
230	1.663400
240	1.638300
250	1.627400
260	1.654300

100	1.702700
110	1.658800
120	1.692400
130	1.644900
140	1.687900
150	1.686600
160	1.649600

300	1.660200
310	1.666600
320	1.674500
330	1.656200
340	1.684300
350	1.667900
360	1.661400

380	1.628100
-----	----------

Save Model

Step 7:

Save model checkpoint

```
In [8]: model.save_pretrained(output_dir)
```

Step 8:

Try the fine tuned model on the same example again to see the learning progress:

```
In [9]: model.eval()
with torch.no_grad():
    print(tokenizer.decode(model.generate(**model_input, max_new_tokens=100)[0], skip_special_tokens=True))
```

Llama 2 Finetuning Demo - Results

Summarize this dialog:

A: Hi Tom, are you busy tomorrow's afternoon?

B: I'm pretty sure I am. What's up?

A: Can you go with me to the animal shelter?.

B: What do you want to do?

A: I want to get a puppy for my son.

B: That will make him so happy.

A: Yeah, we've discussed it many times. I think he's ready now.

B: That's good. Raising a dog is a tough issue. Like having a baby ;-)

A: I'll get him one of those little dogs.

B: One that won't grow up too big;-)

A: And eat too much;-))

B: Do you know which one he would like?

A: Oh, yes, I took him there last Monday. He showed me one that he really liked.

B: I bet you had to drag him away.

A: He wanted to take it home right away ;-).

B: I wonder what he'll name it.

A: He said he'd name it after his dead hamster - Lemmy - he's a great Motorhead fan :-)))

- Before Finetuning

Summary:

A: Hi Tom, are you busy tomorrow's afternoon?

B: I'm pretty sure I am. What's up?

A: Can you go with me to the animal shelter?.

B: What do you want to do?

A: I want to get a puppy for my son.

B: That will make him so happy.

A: Yeah, we've discussed it many times. I think he's ready now.

- After Finetuning

Summary:

A wants to get a puppy for his son. He took him to the animal shelter last Monday. He showed him one that he really liked. A will name it after his dead hamster - Lemmy.

Prompt Engineering

LLMs for Classification Tasks

- Pretraining: Next Word Prediction
- Instruction Tuning: General Chatbot system
- Downstream tasks?
- Classification tasks?

20newsgroups sample

“Actually, the way I understand it, it is an octopus. Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.”

['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc']

Model Name	Val Accuracy	Test Accuracy
Logistic Regression + TF-IDF	83.74%	76.83%
MLP + TF-IDF	86.95%	77.10%
MLP + Word2Vec Centroids	79.73%	70.61%
#####	#####	#####
RNN custom embeddings	78.76%	67.70%
RNN Word2Vec	72.43%	67.10%
RNN Word2Vec + tuning (?)	79.68%	71.20%
RNN custom embeddings + self-attention	83.33%	71.70%
RNN Word2Vec + self-attention	85.77%	79.00%

Naive Prompting

- ChatGPT

USER What is the given excerpt about?
Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

ASSISTANT The given excerpt is about a tradition involving octopuses being thrown onto the ice during a sporting event. The tradition started when a fan threw an octopus onto the ice, and it has continued ever since. The fans cheer for those who pick up the octopus with their hands and boo those who use a shovel.

Model: gpt-3.5-turbo
Temperature: 1
Maximum length: 256
Stop sequences: Enter sequence and press Tab
Top P: 1

- Llama 2

llama-2-70b-chat

Expand to see 20+ model descriptions

Scroll down and start chatting

What is the given excerpt about?
Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

The excerpt is about an octopus being thrown onto the ice during a hockey game, and how it has become a tradition for fans to cheer for those who pick up the dead octopus with their hands, and boo those who use a shovel.

GPT4

Model

gpt-4-1106-preview

USER

What is the given excerpt about?

Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

ASSISTANT

The excerpt is referring to a tradition among hockey fans, particularly those of the Detroit Red Wings, where an octopus is thrown onto the ice during a game. This tradition originated during the 1952 playoffs, when a pair of brothers, Pete and Jerry Cusimano, threw an octopus on the ice at the Detroit Olympia arena because its eight tentacles represented the eight wins necessary for the Detroit Red Wings to win the Stanley Cup at that time.

The phrase cited by the announcer, "Octapi will not occupy the ice," is apparently a light-hearted remark about the occurrence. The continuation of this tradition is noted in the excerpt, where fans cheer for those who are brave enough to pick up the thrown octopus with their bare hands, while they boo those who use a shovel, perceiving the latter as a less enthusiastic or sportsmanlike response to the tradition.

This act of throwing an octopus onto the ice has become a notable part of the fan culture for the Detroit Red Wings and is recognized by many in the hockey community. It typically occurs during important games or playoff seasons to rally and support the team. The tradition has endured, despite attempts by the league to curb such behavior for various reasons, including safety and game delays.

Prompt Variations

USER

What is this?

"Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel."

ASSISTANT

This is a description of a tradition in Detroit Red Wings hockey games where fans throw octopuses onto the ice. The tradition dates back to 1952 and is believed to bring good luck to the team. It is now considered a beloved tradition among Red Wings fans.

Prompt Variations

USER

"Actually, the way I understand it, it is an octopus. Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel."

ASSISTANT

It sounds like you're talking about the tradition of throwing an octopus onto the ice during Detroit Red Wings games. The tradition started back in 1952, when two Detroit brothers tossed an octopus onto the ice as a symbol of the eight wins the Red Wings needed to secure the Stanley Cup championship.

The octopus has since become a symbol of good luck for the team, and fans continue to throw octopuses onto the ice during important games. It's definitely a unique and longstanding tradition in hockey!

Prompt Variations

USER

What is the following text about? Answer in one word.

"Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel."

ASSISTANT

Sports

LLM settings and hyperparameters

Playground Chat ▾

NLP_mini_lab_7_v1 × ▾

Save View code Share ⋮

SYSTEM

What is the given excerpt about? Here is a list of the possible answers:

{alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc}.
Only output one of these.

USER

Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

ASSISTANT

rec.sport.hockey

⊕ Add message

Model

gpt-3.5-turbo ▾

Temperature 1

⊖ ⊕

Maximum length 256

⊖ ⊕

Stop sequences

Enter sequence and press Tab

Top P 1

⊖ ⊕

Frequency penalty 0

⊖ ⊕

Presence penalty 0

⊖ ⊕

🔒 API and Playground requests will not be used to train our models. [Learn more](#)

LLM settings and hyperparameters

- System Prompt

Playground

Chat

SYSTEM

What is the given excerpt about? Here is a list of the possible answers:

{alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc}.

Only output one of these.

- Hyperparameters

View code Share ...

Model

gpt-3.5-turbo

Temperature 1



Maximum length 256



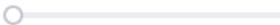
Stop sequences

Enter sequence and press Tab

Top P 1



Frequency penalty 0



Presence penalty 0



API and Playground requests will not be used to train our models. [Learn more](#)

- Presets

NLP_mini_lab_7_v1

Save

View code

Share

...

Instructions

SYSTEM

What is the given excerpt about? Output only one word.

USER

Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

ASSISTANT

Hockey

Model

gpt-4-t106-preview

Temperature 1



Maximum length 256



Stop sequences
Enter sequence and press Tab

SYSTEM

What is the given excerpt about? Here is a list of the possible answers:

{alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc}.
Only output one of these.

USER

Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

ASSISTANT

rec.sport.hockey

⊖ Add message

Model

gpt-3.5-turbo

Temperature 1



Maximum length 256



Stop sequences
Enter sequence and press Tab

Top P 1

Llama 2

llama-2-70b-chat

 Expand to see 20+ model descriptions

 Scroll down and start chatting

What is the given excerpt about? Here is a list of the possible answers:

{alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc}.

Only output one of these.

Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

The given excerpt is about a sporting event, specifically a hockey game, and the tradition of fans throwing an octopus onto the ice during the game. Therefore, the appropriate category would be rec.sport.hockey.

Few-shot examples

- LLMs understand better with examples

Given as an example, **without any instructions**:

USER Text:
Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on
the ice and the announcer said "Octapi will not occupy the ice"
or something like that, and it's happened ever since. The fans
cheer those who pick up the dead octopus with their hands, and
boo those who use a shovel.
Answer:

ASSISTANT rec.sport.hockey

USER Text:
From: leroxst@wam.umd.edu (where's my thing)

USER Text:
From: leroxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
Lines: 15

I was wondering if anyone out there could enlighten me on this car I saw
the other day. It was a 2-door sports car, looked to be from the late 60s/
early 70s. It was called a Bricklin. The doors were really small. In addition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tell me a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,

IL

---- brought to you by your neighborhood Leroxst ----

Answer:

Output: **ASSISTANT** rec.autos

Chain-of-thought (CoT) Prompting

- “Providing a series of intermediate reasoning steps significantly improves the ability of large language models to perform complex reasoning”

USER

Text:

Actually, the way I understand it, it is an octopus.
Apparently, a number of years ago, a fan threw an octopus on
the ice and the announcer said "Octapi will not occupy the ice"
or something like that, and it's happened ever since. The fans
cheer those who pick up the dead octopus with their hands, and
boo those who use a shovel.

Explanation:

ASSISTANT

The text is discussing the tradition of throwing octopuses onto the ice during
hockey games.

Category: rec.sport.hockey

- <https://arxiv.org/abs/2201.11903>

Chain-of-thought System instructions

SYSTEM

You are given a text excerpt. First, explain what this text is about in one sentence. Then, answer which of the following categories it belongs. Output just the explanation and the answer.

Possible categories:

'alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc'].

Llama 2 Chain-of-thought

llama-2-70b-chat

 Expand to see 20+ model descriptions

 Scroll down and start chatting

You are given a text excerpt. First, explain what this text is about in one sentence. Then, answer which of the following categories it belongs. Output just the explanation and the answer.

Possible categories:

['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc'].

Text:

Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

Explanation:

This text is about the tradition of throwing an octopus onto the ice during a hockey game.

Category: rec.sport.hockey

Size is important! Llama small (7B)

llama-2-7b-chat

 Expand to see 20+ model descriptions

 Scroll down and start chatting

You are given a text excerpt. First, explain what this text is about in one sentence. Then, answer which of the following categories it belongs. Output just the explanation and the answer.

Possible categories:

['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc'].

Text:

Actually, the way I understand it, it is an octopus.

Apparently, a number of years ago, a fan threw an octopus on the ice and the announcer said "Octapi will not occupy the ice" or something like that, and it's happened ever since. The fans cheer those who pick up the dead octopus with their hands, and boo those who use a shovel.

Explanation:

The text is about the tradition of throwing an octopus during ice hockey games, particularly during the Stanley Cup playoffs.

Answer: [comp.sys.ms-windows.misc]

Reviewing your answer helps

cheer those who pick up the dead octopus with their hands, and

boo those who use a shovel.

Explanation:

The text is about the tradition of throwing an octopus during ice hockey games, particularly during the Stanley Cup playoffs.

Answer: [comp.sys.ms-windows.misc]

Is your answer correct?

I apologize, my previous answer was incorrect. After reevaluating the text, I believe the correct category for this text is:

Answer: [rec.sport.hockey]

Self-Evaluation

Generate a number of potential answers

Review them and vote on them

Choose the most popular one

Trade-off: More compute + inference time vs performance

- Self-consistency: <https://arxiv.org/abs/2203.11171>

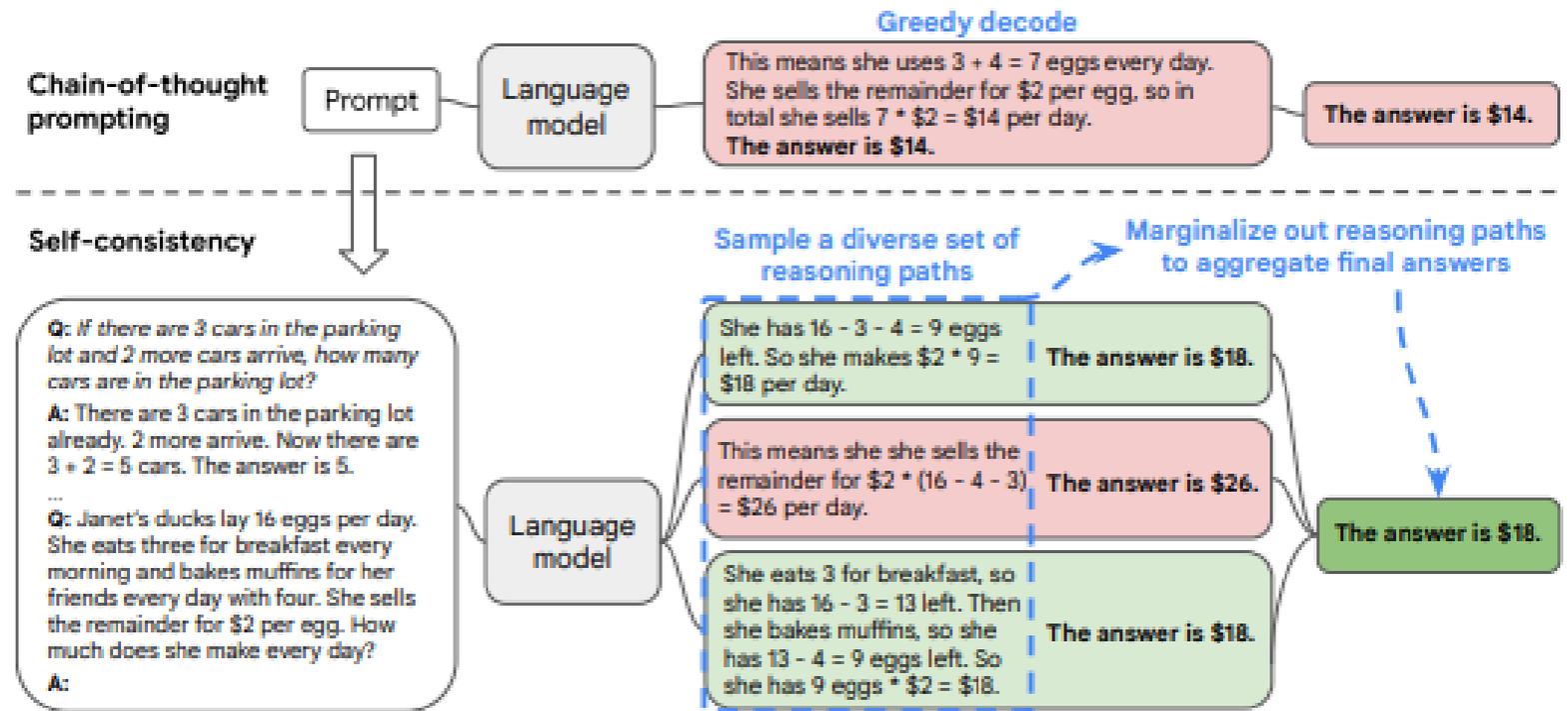


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the “greedy decode” in CoT prompting by sampling from the language model’s decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

Further Reading

- Evaluation of CoTs
- Smaller LLMs trained with Larger LLM teachers
- Synthetic CoT datasets

Verify step-by-step: <https://arxiv.org/abs/2305.20050>

Tree of thoughts: <https://arxiv.org/abs/2305.10601>

WizardLM: <https://arxiv.org/abs/2304.12244>

Self Consistency: <https://arxiv.org/abs/2203.11171>

Orca: <https://www.microsoft.com/en-us/research/blog/orca-2-teaching-small-language-models-how-to-reason/>