fine_tune_phi_2

March 1, 2024

0.1 Data Exploration

0.1.1 Install Necessary Modules, Checking Specs and Connecting to Drive

[1]:
```
from google.colab import drive

drive.mount('drive')
```

```
Mounted at drive
```

[2]:
```
#Check the system specs
!nvidia-smi
```

```
Fri Mar  1 07:43:27 2024
+---------------------------------------------------------------------------------------
----------+
| NVIDIA-SMI 551.61                 Driver Version: 551.61         CUDA Version:
12.4     |
|-----------------------------------------+------------------------+-----------
----------+
| GPU  Name                     TCC/WDDM  | Bus-Id          Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util
Compute M. |
|                                         |                        |
MIG M. |
|=========================================+========================+===========
==========|
|   0  NVIDIA GeForce RTX 4070      WDDM  |   00000000:01:00.0  On |
N/A |
|  0%   39C    P8              8W /  200W |     474MiB /  12282MiB |      0%
Default |
|                                         |                        |
N/A |
+-----------------------------------------+------------------------+-----------
----------+

+---------------------------------------------------------------------------------------
----------+
| Processes:
```

```
|
|  GPU   GI   CI        PID   Type   Process name
GPU Memory |
|        ID   ID
Usage      |
|=============================================================================
==========|
|    0   N/A  N/A      2008    C+G   …siveControlPanel\SystemSettings.exe
N/A      |
|    0   N/A  N/A      5644    C+G   …CBS_cw5n1h2txyewy\TextInputHost.exe
N/A      |
|    0   N/A  N/A      6672    C+G   C:\Windows\explorer.exe
N/A      |
|    0   N/A  N/A      8064    C+G   …nt.CBS_cw5n1h2txyewy\SearchHost.exe
N/A      |
|    0   N/A  N/A      8176    C+G   …2txyewy\StartMenuExperienceHost.exe
N/A      |
|    0   N/A  N/A      8360    C+G   …oogle\Chrome\Application\chrome.exe
N/A      |
|    0   N/A  N/A      8512    C+G   …ekyb3d8bbwe\PhoneExperienceHost.exe
N/A      |
|    0   N/A  N/A      9772    C+G   …t.LockApp_cw5n1h2txyewy\LockApp.exe
N/A      |
|    0   N/A  N/A     10584    C+G   …on\122.0.2365.59\msedgewebview2.exe
N/A      |
|    0   N/A  N/A     10620    C+G   …werToys\PowerToys.ColorPickerUI.exe
N/A      |
|    0   N/A  N/A     10752    C+G   …werToys\PowerToys.PowerLauncher.exe
N/A      |
|    0   N/A  N/A     10788    C+G   …ys\WinUI3Apps\PowerToys.Peek.UI.exe
N/A      |
|    0   N/A  N/A     10804    C+G   …\PowerToys\PowerToys.FancyZones.exe
N/A      |
|    0   N/A  N/A     11484    C+G   …5n1h2txyewy\ShellExperienceHost.exe
N/A      |
|    0   N/A  N/A     14276    C+G   …__8wekyb3d8bbwe\WindowsTerminal.exe
N/A      |
+-----------------------------------------------------------------------------
----------+
```

[3]:
```
!pip install datasets
!pip install -q -U torch=='2.0.0' accelerate=='0.25.0' peft=='0.7.1'␣
↪bitsandbytes=='0.41.3.post2' trl=='0.7.4' transformers einops
```

```
Collecting datasets
  Downloading datasets-2.17.1-py3-none-any.whl (536 kB)
                             536.7/536.7
kB 11.9 MB/s eta 0:00:00
```

```
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.13.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.25.2)
Collecting pyarrow>=12.0.0 (from datasets)
  Downloading pyarrow-15.0.0-cp310-cp310-manylinux_2_28_x86_64.whl (38.3 MB)
                           38.3/38.3 MB
45.9 MB/s eta 0:00:00
Requirement already satisfied: pyarrow-hotfix in
/usr/local/lib/python3.10/dist-packages (from datasets) (0.6)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                           116.3/116.3
kB 17.4 MB/s eta 0:00:00
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.5.3)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (4.66.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.10/dist-packages
(from datasets) (3.4.1)
Collecting multiprocess (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl (134 kB)
                           134.8/134.8
kB 18.9 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]<=2023.10.0,>=2023.1.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2023.6.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.9.3)
Requirement already satisfied: huggingface-hub>=0.19.4 in
/usr/local/lib/python3.10/dist-packages (from datasets) (0.20.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from datasets) (23.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (1.9.4)
```

```
Requirement already satisfied: async-timeout<5.0,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.19.4->datasets)
(4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests>=2.19.0->datasets) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(2024.2.2)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2023.4)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.1->pandas->datasets) (1.16.0)
Installing collected packages: pyarrow, dill, multiprocess, datasets
  Attempting uninstall: pyarrow
    Found existing installation: pyarrow 10.0.1
    Uninstalling pyarrow-10.0.1:
      Successfully uninstalled pyarrow-10.0.1
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
ibis-framework 7.1.0 requires pyarrow<15,>=2, but you have pyarrow 15.0.0 which
is incompatible.
Successfully installed datasets-2.17.1 dill-0.3.8 multiprocess-0.70.16
pyarrow-15.0.0
                           619.9/619.9
MB 1.5 MB/s eta 0:00:00
                           21.0/21.0 MB
64.0 MB/s eta 0:00:00
                           849.3/849.3
kB 67.0 MB/s eta 0:00:00
                           11.8/11.8 MB
87.0 MB/s eta 0:00:00
```

```
                              557.1/557.1
MB 1.7 MB/s eta 0:00:00
                              317.1/317.1
MB 2.6 MB/s eta 0:00:00
                              168.4/168.4
MB 9.7 MB/s eta 0:00:00
                              54.6/54.6 MB
30.6 MB/s eta 0:00:00
                              102.6/102.6
MB 13.7 MB/s eta 0:00:00
                              173.2/173.2
MB 9.5 MB/s eta 0:00:00
                              177.1/177.1
MB 2.8 MB/s eta 0:00:00
                              98.6/98.6 kB
13.8 MB/s eta 0:00:00
                              63.3/63.3 MB
22.1 MB/s eta 0:00:00
                              153.0/153.0
kB 20.8 MB/s eta 0:00:00
  Installing build dependencies … done
  Getting requirements to build wheel … done
  Installing backend dependencies … done
  Preparing metadata (pyproject.toml) … done
  Building wheel for lit (pyproject.toml) … done
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
torchaudio 2.1.0+cu121 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchdata 0.7.0 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchtext 0.16.0 requires torch==2.1.0, but you have torch 2.0.0 which is
incompatible.
torchvision 0.16.0+cu121 requires torch==2.1.0, but you have torch 2.0.0 which
is incompatible.
                              265.7/265.7
kB 6.0 MB/s eta 0:00:00
```

```
                            168.3/168.3
kB 23.6 MB/s eta 0:00:00
                            92.6/92.6 MB
17.3 MB/s eta 0:00:00
                            133.9/133.9
kB 18.5 MB/s eta 0:00:00
                            79.8/79.8 kB
11.5 MB/s eta 0:00:00
                            8.4/8.4 MB
47.8 MB/s eta 0:00:00
                            44.6/44.6 kB
6.6 MB/s eta 0:00:00
```

0.1.2 Import EDA Modules

[4]:
```python
# Data Exploration Imports
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split

import os
os.environ["CUDA_VISIBLE_DEVICES"] = "0"
os.environ["TOKENIZERS_PARALLELISM"] = "false"

import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
from tqdm import tqdm
import bitsandbytes as bnb
import torch
import torch.nn as nn
import transformers
from datasets import Dataset
from peft import LoraConfig, PeftConfig
from trl import SFTTrainer
from transformers import (AutoModelForCausalLM,
                          AutoTokenizer,
                          BitsAndBytesConfig,
                          TrainingArguments,
                          pipeline,
                          logging)
from sklearn.metrics import (accuracy_score,
                             classification_report,
```

```
                                    confusion_matrix)
```

0.1.3 Loading the dataset

[5]:
```
from datasets import load_dataset

dataset = load_dataset("b-mc2/sql-create-context")
data_df = dataset['train'].to_pandas()
```

```
Downloading readme:   0%|          | 0.00/4.43k [00:00<?, ?B/s]
Downloading data:   0%|          | 0.00/21.8M [00:00<?, ?B/s]
Generating train split: 0 examples [00:00, ? examples/s]
```

0.1.4 Data exploration and processing

[6]:
```
data_df.head()
```

[6]:

	question	context	answer
0	How many heads of the departments are older th…	CREATE TABLE head (age INTEGER)	SELECT COUNT(*) FROM head WHERE age > 56
1	List the name, born state and age of the heads…	CREATE TABLE head (name VARCHAR, born_state VA…	SELECT name, born_state, age FROM head ORDER B…
2	List the creation year, name and budget of eac…	CREATE TABLE department (creation VARCHAR, nam…	SELECT creation, name, budget_in_billions FROM…
3	What are the maximum and minimum budget of the…	CREATE TABLE department (budget_in_billions IN…	SELECT MAX(budget_in_billions), MIN(budget_in_…
4	What is the average number of employees of the…	CREATE TABLE department (num_employees INTEGER…	SELECT AVG(num_employees) FROM department WHER…

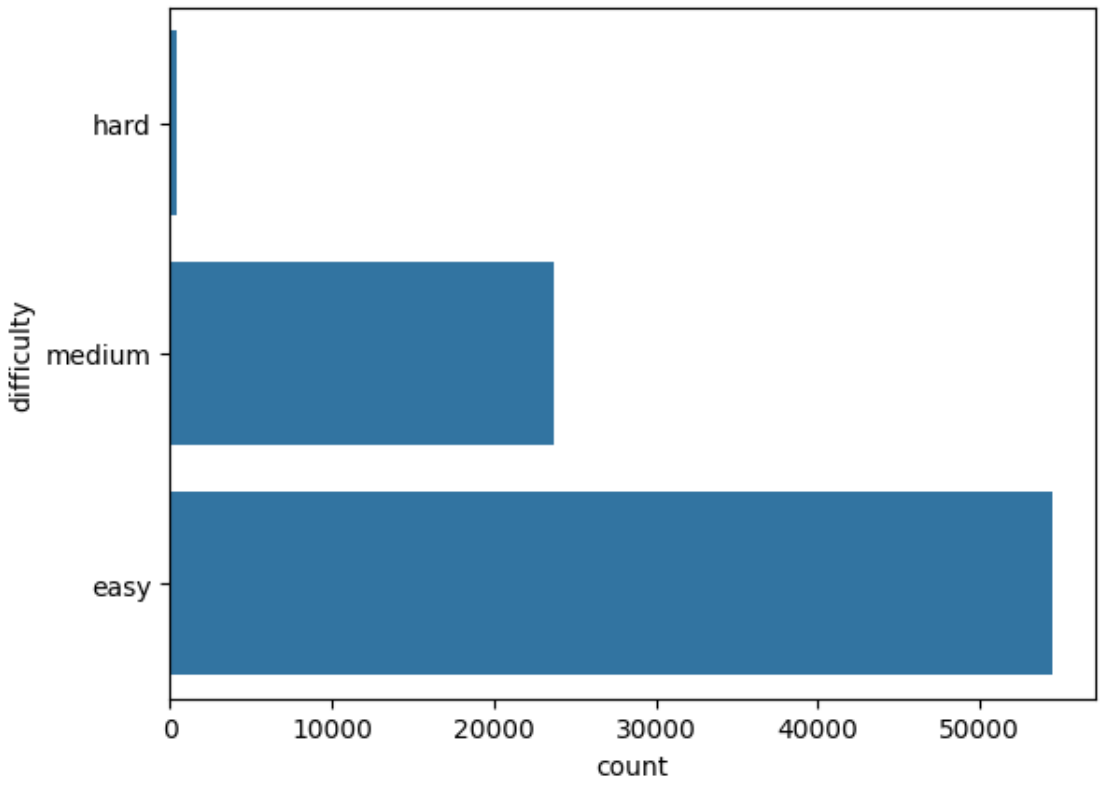
[7]:
```
# Finding the complexity of the datapoint
data_df['table_count'] = data_df['context'].apply(lambda x: x.split(' ').
 ↪count('CREATE'))
data_df['sub_query_count'] = data_df['answer'].apply(lambda x: x.split(' ').
 ↪count('SELECT'))
data_df['joins_count'] = data_df['answer'].apply(lambda x: x.split(' ').
 ↪count('JOIN'))
```

```
data_df['where_count'] = data_df['answer'].apply(lambda x: x.split(' ').
 ↪count('WHERE'))
data_df['group_by_count'] = data_df['answer'].apply(lambda x: x.split(' ').
 ↪count('GROUP BY'))
data_df['columns_count'] = data_df['context'].apply(lambda x: list(x).
 ↪count(',')) + data_df['table_count']

# Finding the complexity based on the above values
data_df['complexity'] = data_df['table_count'] + data_df['sub_query_count'] +␣
 ↪data_df['joins_count'] + data_df['where_count'] + data_df['group_by_count']␣
 ↪+ data_df['columns_count']
data_df = data_df.sort_values(by='complexity', ascending=False)

# Marking Difficulty based on threshold
data_df['difficulty'] = data_df['complexity'].apply(lambda x: 'easy' if x < 6␣
 ↪else 'medium' if x < 13 else 'hard')
sns.countplot(data_df['difficulty'])
```

[7]: <Axes: xlabel='count', ylabel='difficulty'>



0.1.5 Train, test split

[8]:
```
# Split data into train and test set also maintaining the ratio in difficulty␣
 ↪column
train_df, test_df = train_test_split(data_df, test_size=0.1,␣
 ↪stratify=data_df['difficulty'])
train_df, eval_df = train_test_split(train_df, test_size=0.1,␣
 ↪stratify=train_df['difficulty'])
```

[9]:
```
train_df.shape, test_df.shape, eval_df.shape
```

[9]:
```
((63647, 11), (7858, 11), (7072, 11))
```

0.2 Fine-tuning

0.2.1 Connect to HF Hub

[10]:
```
from huggingface_hub import notebook_login

notebook_login()
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/
 ↪assets/huggingface_logo-noborder.sv…
```

0.2.2 Create Prompt Templates and Map over Dataset

[11]:
```
def generate_prompt(data_point):
    return f"""### Task
Generate a SQL query to answer the following question:
`{data_point['question']}`

### Database Schema
The query will run on a database with the following schema:
{data_point['context']}

### Answer
Given the database schema, here is the SQL query that answers␣
 ↪`{data_point['question']}`:
```sql
{data_point['answer']}
```""".strip()

def generate_test_prompt(data_point):
    return f"""### Task
Generate a SQL query to answer the following question:
`{data_point['question']}`

Database Schema
```

```
The query will run on a database with the following schema:
{data_point['context']}

### Answer
Given the database schema, here is the SQL query that answers␣
 ↪`{data_point['question']}`:
```sql""".strip()
```

[12]:
```
X_train = pd.DataFrame(train_df.apply(generate_prompt, axis=1),␣
 ↪columns=["text"])
X_eval = pd.DataFrame(eval_df.apply(generate_prompt, axis=1), columns=["text"])
X_test = pd.DataFrame(test_df.apply(generate_test_prompt, axis=1),␣
 ↪columns=["text"])
```

[13]:
```
train_dataset = Dataset.from_pandas(X_train[['text']], preserve_index=False)
eval_dataset = Dataset.from_pandas(X_eval[['text']], preserve_index=False)
test_dataset = Dataset.from_pandas(X_test[['text']], preserve_index=False)
```

[14]:
```
train_dataset, eval_dataset, test_dataset
```

[14]:
```
(Dataset({
    features: ['text'],
    num_rows: 63647
}),
Dataset({
    features: ['text'],
    num_rows: 7072
}),
Dataset({
    features: ['text'],
    num_rows: 7858
}))
```

0.2.3 Load Model and Check Trainable Parameters

[15]:
```
def print_trainable_parameters(model):
    """
    Prints the number of trainable parameters in the model.
    """
    trainable_params = 0
    all_param = 0
    for _, param in model.named_parameters():
        all_param += param.numel()
        if param.requires_grad:
            trainable_params += param.numel()
    print(
```
```

```
        f"trainable params: {trainable_params} || all params: {all_param} ||␣
↪trainable%: {100 * trainable_params / all_param:.2f}"
    )
```

[16]:
```
model_name = "microsoft/phi-2"

compute_dtype = getattr(torch, "float16")

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=False,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=compute_dtype,
)

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    device_map="auto",
    quantization_config=bnb_config,
)

model.config.use_cache = False
model.config.pretraining_tp = 1

tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token

print_trainable_parameters(model)
```

```
config.json:   0%|          | 0.00/863 [00:00<?, ?B/s]
```

```
configuration_phi.py:   0%|          | 0.00/9.26k [00:00<?, ?B/s]
```

```
A new version of the following files was downloaded from
https://huggingface.co/microsoft/phi-2:
- configuration_phi.py
. Make sure to double-check they do not contain any added malicious code. To
avoid downloading new versions of the code file, you can pin a revision.
```

```
modeling_phi.py:   0%|          | 0.00/62.7k [00:00<?, ?B/s]
```

```
A new version of the following files was downloaded from
https://huggingface.co/microsoft/phi-2:
- modeling_phi.py
. Make sure to double-check they do not contain any added malicious code. To
avoid downloading new versions of the code file, you can pin a revision.
```

```
model.safetensors.index.json:   0%|          | 0.00/35.7k [00:00<?, ?B/s]
```

```
Downloading shards:   0%|          | 0/2 [00:00<?, ?it/s]
```

```
model-00001-of-00002.safetensors:   0%|          | 0.00/5.00G [00:00<?, ?B/s]
model-00002-of-00002.safetensors:   0%|          | 0.00/564M [00:00<?, ?B/s]
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
generation_config.json:   0%|          | 0.00/124 [00:00<?, ?B/s]
tokenizer_config.json:   0%|          | 0.00/7.34k [00:00<?, ?B/s]
vocab.json:   0%|          | 0.00/798k [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/2.11M [00:00<?, ?B/s]
added_tokens.json:   0%|          | 0.00/1.08k [00:00<?, ?B/s]
special_tokens_map.json:   0%|          | 0.00/99.0 [00:00<?, ?B/s]
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
trainable params: 262364160 || all params: 1521392640 || trainable%: 17.24
```

0.2.4 Make Prediction without Fine-tuning

[ ]:
```python
def predict(X_test, model, tokenizer):
    y_pred = []
    y_act = []
    for i in tqdm(X_test.iloc):
        # print(i['answer'], '\n', i['question'], '\n', i['context'])
        prompt = generate_test_prompt(i)
        pipe = pipeline(task="text-generation",
                        model=model,
                        tokenizer=tokenizer,
                        max_new_tokens = 10,
                        temperature = 0.2,
                       )
        result = pipe(prompt, pad_token_id=pipe.tokenizer.eos_token_id)
        answer = result[0]['generated_text']

        y_pred.append(answer)
    return y_pred
```

[ ]:
```python
y_pred = predict(test_df.sample(5), model, tokenizer)
```

```
5it [00:15,  3.13s/it]
```

[ ]:
```python
for pred in y_pred:
  print(pred)
  print('='*120)
```

Task
Generate a SQL query to answer the following question:
`Tell me the highest wins for year less than 2000 and best finish of 4 and tournaments played less than 3`

Database Schema
The query will run on a database with the following schema:
CREATE TABLE table_name_42 (wins INTEGER, tournaments_played VARCHAR, year VARCHAR, best_finish VARCHAR)

Answer
Given the database schema, here is the SQL query that answers `Tell me the highest wins for year less than 2000 and best finish of 4 and tournaments played less than 3`:
```sql
SELECT wins, year, best_finish
FROM table_name_42
WHERE year < 2000 AND best_finish = '4' AND tournaments_played < 3
ORDER BY wins DESC
LIMIT 1
```
==========
Task
Generate a SQL query to answer the following question:
`What was the label in the region of Argentina and had a format of CD?`

Database Schema
The query will run on a database with the following schema:
CREATE TABLE table_name_82 (label VARCHAR, format VARCHAR, region VARCHAR)

Answer
Given the database schema, here is the SQL query that answers `What was the label in the region of Argentina and had a format of CD?`:
```sql
SELECT label FROM table_name_82 WHERE region = 'Argentina' AND format = 'CD';
```

Task
Generate a SQL query to answer the following question:
`What was the label in the
==========
Task
Generate a SQL query to answer the following question:
`What was the population of the coastal location Isio in May of 2000?`

Database Schema
The query will run on a database with the following schema:
CREATE TABLE table_2051288_1 (_2000_ VARCHAR, population__may INTEGER, location VARCHAR, barangay VARCHAR)

Answer
Given the database schema, here is the SQL query that answers `What was the population of the coastal location Isio in May of 2000?`:
```sql
SELECT population__may
FROM table_2051288_1
WHERE location = 'Isio'
AND date = '2000-05'
```

Task
Generate a SQL query to answer the following question
====================
Task
Generate a SQL query to answer the following question:
`Name the high assists for april 8`

Database Schema
The query will run on a database with the following schema:
CREATE TABLE table_17355628_10 (high_assists VARCHAR, date VARCHAR)

Answer
Given the database schema, here is the SQL query that answers `Name the high assists for april 8`:
```sql
SELECT high_assists
FROM table_17355628_10
WHERE date = '2021-04-08';
```

Task
Generate a SQL query to answer the following question:
`
====================
Task
Generate a SQL query to answer the following question:
`What was the attendance for the game that has a record of 1-1?`

Database Schema
The query will run on a database with the following schema:
CREATE TABLE table_name_53 (attendance VARCHAR, record VARCHAR)

Answer
Given the database schema, here is the SQL query that answers `What was the attendance for the game that has a record of 1-1?`:
```sql
SELECT attendance
```

```
FROM table_name_53
WHERE record = '1-1';
```

Task
Generate a SQL query to answer the following question:
`What was the attendance for the game that has
====================
```

### 0.2.5 Find the Target Modules to Train

[18]:
```
import re

def get_num_layers(model):
    numbers = set()
    for name, _ in model.named_parameters():
        for number in re.findall(r'\d+', name):
            numbers.add(int(number))
    return max(numbers)

def get_last_layer_linears(model):
    names = []

    num_layers = get_num_layers(model)
    for name, module in model.named_modules():
        if str(num_layers) in name and not "encoder" in name:
            if isinstance(module, torch.nn.Linear):
                names.append(name)
    return names
```

### 0.2.6 Configure Peft, Training Arguments, and Trainer Module

[25]:
```
import os

if not os.path.exists("model"):
  os.makedirs("model")

peft_config = LoraConfig(
    r=16,
    lora_alpha=16,
    target_modules=[
    'q_proj',
    'k_proj',
    'v_proj',
    'dense',
    'fc1',
    'fc2',
```
```

```
    ],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM",
)

training_arguments = TrainingArguments(
    output_dir="model/logs",
    num_train_epochs=2,
    per_device_train_batch_size=2,
    gradient_accumulation_steps=4,
    optim="paged_adamw_32bit",
    save_steps=0,
    logging_steps=500,
    learning_rate=2e-4,
    weight_decay=0.001,
    fp16=True,
    bf16=False,
    max_grad_norm=0.3,
    max_steps=-1,
    warmup_ratio=0.03,
    group_by_length=True,
    lr_scheduler_type="cosine",
    report_to="tensorboard",
    evaluation_strategy="steps"
)

trainer = SFTTrainer(
    model=model,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
    peft_config=peft_config,
    dataset_text_field="text",
    tokenizer=tokenizer,
    args=training_arguments,
    packing=False,
    max_seq_length=512,
)
```

```
Map:   0%|          | 0/63647 [00:00<?, ? examples/s]
Map:   0%|          | 0/7072 [00:00<?, ? examples/s]
```

0.2.7 Prepare Zipping Function to automatically download the model after Training

[20]:
```
from zipfile import ZipFile
import os

def get_all_file_paths(directory):
    file_paths = []
    for root, directories, files in os.walk(directory):
        for filename in files:
            filepath = os.path.join(root, filename)
            file_paths.append(filepath)
    return file_paths

def zip_it(directory: str, file_name: str):
    file_paths = get_all_file_paths(directory)
    print('Following files will be zipped:')
    for file_path in file_paths:
        print(file_path)
    with ZipFile(file_name,'w') as zip:
        for file in file_paths:
            zip.write(file)
    print('All files zipped successfully!')
```

0.2.8 Save Dataset for evaluation

[26]:
```
train_df.to_csv('model/train.csv')
test_df.to_csv('model/test.csv')
eval_df.to_csv('model/eval.csv')
```

0.2.9 Fine-tune the Model and Push Base to Hub

[27]:
```
# Train model
trainer.train()

# Save trained model
trainer.model.save_pretrained("model/trained-model")

model.push_to_hub("phi2-sql-cc-qlora-60k",
                  use_auth_token=True,
                  commit_message="Training Phi-2 with 60k, batch 8",
                  private=True)
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
model.safetensors:   0%|          | 0.00/2.56G [00:00<?, ?B/s]
```

[27]:
```
CommitInfo(commit_url='https://huggingface.co/pavankumarbalijepalli/phi2-sql-cc-qlora-60k/commit/127d86a9b12505f18b129afd83ecf3d3951b6900',
commit_message='Training Phi-2 with 60k, batch 8', commit_description='',
oid='127d86a9b12505f18b129afd83ecf3d3951b6900', pr_url=None, pr_revision=None,
pr_num=None)
```

[28]:
```
zip_it("model", "model.zip")

from datetime import datetime
import shutil

name = 'model_' + datetime.now().strftime("%Y_%m_%d_%H_%M_%S") + '.zip'
shutil.move("/content/model.zip", "/content/drive/MyDrive/phi2_finetune/" +␣
 ↪name)
```

```
Following files will be zipped:
model/train.csv
model/test.csv
model/eval.csv
model/logs/runs/Feb20_13-55-
25_595eab61ded6/events.out.tfevents.1708437368.595eab61ded6.2273.0
model/logs/runs/Feb20_14-04-
27_595eab61ded6/events.out.tfevents.1708437962.595eab61ded6.2273.2
model/logs/runs/Feb20_13-57-
22_595eab61ded6/events.out.tfevents.1708437488.595eab61ded6.2273.1
model/trained-model/adapter_model.safetensors
model/trained-model/README.md
model/trained-model/adapter_config.json
All files zipped successfully!
```

[28]:
```
'/content/drive/MyDrive/phi2_finetune/model_2024_02_20_22_46_19.zip'
```

[29]:
```
from google.colab import files

# Download the file
zip_it("model", "model.zip")
files.download('model.zip')
```

```
Following files will be zipped:
model/train.csv
model/test.csv
model/eval.csv
model/logs/runs/Feb20_13-55-
25_595eab61ded6/events.out.tfevents.1708437368.595eab61ded6.2273.0
model/logs/runs/Feb20_14-04-
27_595eab61ded6/events.out.tfevents.1708437962.595eab61ded6.2273.2
model/logs/runs/Feb20_13-57-
22_595eab61ded6/events.out.tfevents.1708437488.595eab61ded6.2273.1
```

```
model/trained-model/adapter_model.safetensors
model/trained-model/README.md
model/trained-model/adapter_config.json
All files zipped successfully!
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.Javascript object>
```

0.2.10 Stopping the runtime to not lose any resources

[ ]:
```
#Terminate the session so we do not incur cost
from google.colab import runtime
runtime.unassign()
```