

DAssignment2b (3)

April 20, 2025

1 Deep Learning Assignment 2 - Part 2

1.1 GPT-2 Fine-Tuning for English Lyrics Generation by v jitesh kumar 160122737199

1.1.1 Summary:

- Fine-tuned GPT2-Medium (345M parameters), not GPT2-Small.
- Dataset manually curated using publicly available English song lyrics (50 Cent sample).
- Training performed for 5 epochs with monitoring of training loss.
- After fine-tuning, sample lyrics were generated successfully using different prompts.

1.2 Fine-tuning GPT-2 for English Lyrics Generation

2 Part 2b of assignment: GPT-2 Fine-Tuning on English Lyrics Dataset

2.1 Step 1: Install Required Libraries

```
[ ]: # Install Huggingface Transformers and Datasets libraries
!pip install transformers datasets

# Also installed Accelerate for faster training (optional but useful)
!pip install accelerate
```

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.51.3)

Collecting datasets

Downloading datasets-3.5.0-py3-none-any.whl.metadata (19 kB)

Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)

Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.2)

Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)

Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)

Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-

packages (from transformers) (6.0.2)
 Requirement already satisfied: regex!=2019.12.17 in
 /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
 Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
 packages (from transformers) (2.32.3)
 Requirement already satisfied: tokenizers<0.22,>=0.21 in
 /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
 Requirement already satisfied: safetensors>=0.4.3 in
 /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
 Requirement already satisfied: tqdm=4.27 in /usr/local/lib/python3.11/dist-
 packages (from transformers) (4.67.1)
 Requirement already satisfied: pyarrow>=15.0.0 in
 /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
 Collecting dill<0.3.9,>=0.3.0 (from datasets)
 Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
 Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
 (from datasets) (2.2.2)
 Collecting xxhash (from datasets)
 Downloading
 xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
 (12 kB)
 Collecting multiprocessing<0.70.17 (from datasets)
 Downloading multiprocessing-0.70.16-py311-none-any.whl.metadata (7.2 kB)
 Collecting fsspec<=2024.12.0,>=2023.1.0 (from
 fsspec[http]<=2024.12.0,>=2023.1.0->datasets)
 Downloading fsspec-2024.12.0-py3-none-any.whl.metadata (11 kB)
 Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
 packages (from datasets) (3.11.15)
 Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
 Requirement already satisfied: aiosignal>=1.1.2 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
 Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
 packages (from aiohttp->datasets) (25.3.0)
 Requirement already satisfied: frozenlist>=1.1.1 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
 Requirement already satisfied: multidict<7.0,>=4.5 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.4.3)
 Requirement already satisfied: propcache>=0.2.0 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
 Requirement already satisfied: yarl<2.0,>=1.17.0 in
 /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.19.0)
 Requirement already satisfied: typing-extensions>=3.7.4.3 in
 /usr/local/lib/python3.11/dist-packages (from huggingface-
 hub<1.0,>=0.30.0->transformers) (4.13.2)
 Requirement already satisfied: charset-normalizer<4,>=2 in
 /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
 Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-

```

packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.5.0-py3-none-any.whl (491 kB)
      491.2/491.2 kB
15.7 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
      116.3/116.3 kB
10.7 MB/s eta 0:00:00
Downloading fsspec-2024.12.0-py3-none-any.whl (183 kB)
      183.9/183.9 kB
15.2 MB/s eta 0:00:00
Downloading multiprocessing-0.70.16-py311-none-any.whl (143 kB)
      143.5/143.5 kB
8.6 MB/s eta 0:00:00
Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
      194.8/194.8 kB
16.4 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocessing,
datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2

```

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.

gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2024.12.0 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12 12.5.3.2 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12 12.5.82 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12 12.5.82 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-runtime-cu12 12.5.82 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cudnn-cu12 9.3.0.75 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cufft-cu12 11.2.3.61 which is incompatible.

torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-curand-cu12 10.3.6.82 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cusolver-cu12 11.6.3.83 which is incompatible.

torch 2.6.0+cu124 requires nvidia-cuspars-cu12==12.3.1.170; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuspars-cu12 12.5.1.3 which is incompatible.

torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64",⁴ but you have nvidia-nvjitlink-cu12 12.5.82 which is incompatible.

Successfully installed datasets-3.5.0 dill-0.3.8 fsspec-2024.12.0

multiprocess-0.70.16 xxhash-3.5.0
Requirement already satisfied: accelerate in /usr/local/lib/python3.11/dist-packages (1.5.2)
Requirement already satisfied: numpy<3.0.0,>=1.17 in /usr/local/lib/python3.11/dist-packages (from accelerate) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from accelerate) (24.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages (from accelerate) (6.0.2)
Requirement already satisfied: torch>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from accelerate) (2.6.0+cu124)
Requirement already satisfied: huggingface-hub>=0.21.0 in /usr/local/lib/python3.11/dist-packages (from accelerate) (0.30.2)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from accelerate) (0.5.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0->accelerate) (3.18.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0->accelerate) (2024.12.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0->accelerate) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0->accelerate) (4.67.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.21.0->accelerate) (4.13.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.0.0->accelerate)
 Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=2.0.0->accelerate)
 Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=2.0.0->accelerate)
 Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=2.0.0->accelerate)
 Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=2.0.0->accelerate)
 Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)

```

Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=2.0.0->accelerate)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cuspars-cu12==12.3.1.170 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cuspars-cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
(12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.0.0->accelerate)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.0.0->accelerate) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.0.0->accelerate) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from
sympy==1.13.1->torch>=2.0.0->accelerate) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0->accelerate)
(3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests->huggingface-hub>=0.21.0->accelerate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (2025.1.31)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4
MB)
363.4/363.4 MB
3.2 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-

```

```

manylinux2014_x86_64.whl (13.8 MB)
      13.8/13.8 MB
83.2 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (24.6 MB)
      24.6/24.6 MB
92.2 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (883 kB)
      883.7/883.7 kB
44.9 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl
(664.8 MB)
      664.8/664.8 MB
1.7 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl
(211.5 MB)
      211.5/211.5 MB
11.4 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl (56.3 MB)
      56.3/56.3 MB
41.8 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl (127.9 MB)
      127.9/127.9 MB
19.0 MB/s eta 0:00:00
Downloading nvidia_cusparses_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl (207.5 MB)
      207.5/207.5 MB
4.0 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (21.1 MB)
      21.1/21.1 MB
105.7 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-curand-cu12,
nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-
cuda-cupti-cu12, nvidia-cublas-cu12, nvidia-cusparses-cu12, nvidia-cudnn-cu12,
nvidia-cusolver-cu12
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12

```

```

Found existing installation: nvidia-cufft-cu12 11.2.3.61
Uninstalling nvidia-cufft-cu12-11.2.3.61:
  Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
Attempting uninstall: nvidia-cuda-runtime-cu12
Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
  Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
Attempting uninstall: nvidia-cuda-nvrtc-cu12
Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
  Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
Attempting uninstall: nvidia-cuda-cupti-cu12
Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
  Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
Attempting uninstall: nvidia-cublas-cu12
Found existing installation: nvidia-cublas-cu12 12.5.3.2
Uninstalling nvidia-cublas-cu12-12.5.3.2:
  Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
Attempting uninstall: nvidia-cusparse-cu12
Found existing installation: nvidia-cusparse-cu12 12.5.1.3
Uninstalling nvidia-cusparse-cu12-12.5.1.3:
  Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
Attempting uninstall: nvidia-cudnn-cu12
Found existing installation: nvidia-cudnn-cu12 9.3.0.75
Uninstalling nvidia-cudnn-cu12-9.3.0.75:
  Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
Attempting uninstall: nvidia-cusolver-cu12
Found existing installation: nvidia-cusolver-cu12 11.6.3.83
Uninstalling nvidia-cusolver-cu12-11.6.3.83:
  Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
Successfully installed nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-
cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127
nvidia-cudnn-cu12-9.1.0.70 nvidia-cufft-cu12-11.2.1.3 nvidia-curand-
cu12-10.3.5.147 nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-cu12-12.3.1.170
nvidia-nvjitlink-cu12-12.4.127

```

2.2 Step 2: Download and Load English Lyrics Dataset

```

[ ]: # Step 2: Fixing dataset loading properly

import pandas as pd

# Load the CSV again but without assuming headers
lyrics_df = pd.read_csv(filename, header=None)

# Now, the third column (index 2) contains lyrics

```



```

lyrics_list = lyrics_df[2].dropna().tolist()

# Display a sample lyric
print("\nSample lyric:")
print(lyrics_list[0])

```

Sample lyric:

I change places, to prevent catchin' the cases
 Races, in the faces, hall at you laces
 This is a hit, let's see if homicide trace this

The only thing hotter than my flow is the block (inhale and exhale)
 That's why I left this snow biz, and got into show biz
 Let's get this clear, it ain't on till I say it's on, (pause), it's on
 I'm eatin', y'all niggas fastin' like it's Rimadon
 Bowlish way in Lebanon, know 50 the bomb
 I be at the edge of the bar, sippin' a Don
 I keep the bottle just in case, you never know when it's on
 This worries bump, I can't go wrong, my team's too strong
 You want war? I take you to war, now that my money long
 Why you broke? cat's buy the by lines and fantasize
 The way I'm spittin', put TV's in everything I'm sittin'
 While I'm hot to death, I'm gonna say this to all you playa haters
 Y'all should hate the game, not the playas (c'mon)

(Chorus: repeat 2X)

I change places, to prevent catchin' the cases
 Races, in the faces, hall at you laces
 This is a hit, let's see if homicide trace this
 (50 Cent, let's see if homicide trace this)

Everyday is bugged, niggas'll come to a club
 To try to show you they a thug, instead of showing some love
 Now, what you think you chump me, If I let you bump me
 When I'm about to make a mill, faster than you make a G (haha)
 I know I lie, it's a habit, I vow to clean the city like the mayor
 And in the crack game, I'm a franchise player
 Niggas be thinkin' I be out to lunch with mines
 Then in crunch time, I start hittin' 'em hard with punch lines
 You cats got to be sick, to think 50 can't spit
 Better check my batting average, I always make hits
 My flows leave these rap cats ketto (ketto), all across the metro (metro)
 Plus I pack a cannon, up under my marple cannon
 They fake, they look like money, but ain't worth half the cake
 Have me runnin' from Jake, in a GS with bad brakes
 They want to knock me take, for Christ sakes

(50 Cent, let's see if homicide trace this)
I change places, to prevent catchin' the cases
Races, in the faces, hall at you laces
This is a hit, let's see if homicide trace this
(50 Cent, let's see if homicide trace this)

I change places, to prevent catchin' the cases
Races, in the faces, hall at you laces
This is a hit, let's see if homicide trace this
(50 Cent, let's see if homicide trace this)

Yo, son remember them fake playas
Who try to play us at The Shark Club in Vegas
Had them tight linen blazers, and beat up gators
Lookin' like last year's playas, (pause)
Yeah, I could tell they dough was low
When we came through the do'
I copped a case of Cristal, and copped one bottle of Mo
From the looking through face, and the bulge in his waist, he holdin'
(Yeah he's packin', I can see his rack
The one in the middle, he a big man, I dealt with him son)
Yeah, so I expect look like they ain't had a run, since ' 81
They ain't here on a hunt for food,
So they could catch you, some cash, and expensive jewels
I'm gonna crash 'em with this bottle if he move
I ain't the one son, my shit ain't come easy
It won't go easy, believe me

(Chorus: repeat 3X)
I change places, to prevent catchin' the cases
Races, in the faces, hall at you laces
This is a hit, let's see if homicide trace this
(50 Cent, let's see if homicide trace this)

2.3 Step 3: Load GPT-2 Tokenizer and Prepare Dataset

```
[ ]: # Step 3: Load GPT-2 tokenizer and prepare dataset

from transformers import GPT2Tokenizer

# Load GPT-2 Medium tokenizer
tokenizer = GPT2Tokenizer.from_pretrained('gpt2-medium')

# Add special tokens if missing (important for fine-tuning)
special_tokens_dict = {'pad_token': '[PAD]'}
tokenizer.add_special_tokens(special_tokens_dict)
```

```

# Tokenize the lyrics
encodings = tokenizer(lyrics_list, truncation=True, padding=True,
    ↪max_length=256, return_tensors="pt")

# Display tokenized sample
print("\nTokenized input sample (IDs):")
print(encodings['input_ids'][0])

```

```

tokenizer_config.json:  0%|          | 0.00/26.0 [00:00<?, ?B/s]
vocab.json:            0%|          | 0.00/1.04M [00:00<?, ?B/s]
merges.txt:            0%|          | 0.00/456k [00:00<?, ?B/s]
tokenizer.json:        0%|          | 0.00/1.36M [00:00<?, ?B/s]
config.json:          0%|          | 0.00/718 [00:00<?, ?B/s]

```

Tokenized input sample (IDs):

```

tensor([ 40, 1487, 4113,  11,  284, 2948, 4929,  259,   6, 262,
        2663,  198,   49, 2114,   11,  287,  262, 6698,   11, 6899,
         379,  345,  300, 2114,  198, 1212,  318,  257, 2277,   11,
        1309,  338,  766,  611, 19625, 12854,  428,  198,  198,  464,
         691, 1517, 37546,  621,  616,  5202,  318,  262, 2512,  357,
         259,   71,  1000,  290, 21847,  1000,    8,  198, 2504,  338,
        1521,  314,  1364,  428,  6729,  275,  528,   11,  290, 1392,
         656,  905,  275,  528,  198,  5756,  338,  651,  428, 1598,
          11,  340, 18959,  470,  319, 10597,  314,  910,  340,  338,
         319,   11,  357, 32125,  828,  340,  338,  319,  198,   40,
        1101, 4483,  259,  3256,  331,    6,  439,  299, 6950,  292,
        3049,  259,    6,  588,  340,  338, 29542,  324,  261,  198,
       39961, 1836,  835,  287, 16849,   11,  760,  2026,  262, 5194,
         198,   40,  307,  379,  262,  5743,  286,  262, 2318,   11,
         264, 3974,  259,    6,  257,  2094,  198,   40, 1394,  262,
        9294,  655,  287, 1339,   11,  345, 1239,  760,  618,  340,
         338,  319,  198, 1212, 18572, 13852,   11,  314,  460,  470,
         467, 2642,   11,  616, 1074,  338, 1165, 1913,  198, 1639,
         765, 1175,   30,  314, 1011,  345,  284, 1175,   11,  783,
         326,  616, 1637,  890,  198,  5195,  345,  6265,   30, 3797,
         338, 2822,  262,  416, 3951,  290, 23857, 1096,  198,  464,
         835,  314, 1101,  599,  715,  259,  3256, 1234, 3195,  338,
         287, 2279,  314, 1101,  264,  715,  259,    6,  198, 3633,
         314, 1101, 3024,  284, 1918,   11,  314, 1101, 8066,  910,
         428,  284,  477,  345,  711,   64, 6877,  364,  198,   56,
          6,  439,  815, 5465,  262,  983])

```

2.4 Step 4: Load GPT-2 Model and Fine-Tune on Lyrics

```
[ ]: # Step 4: Load GPT-2 model and fine-tune with evaluation

from transformers import GPT2LMHeadModel, Trainer, TrainingArguments
import torch
from torch.utils.data import Dataset, random_split

# Load GPT-2 medium model
model = GPT2LMHeadModel.from_pretrained('gpt2-medium')

# Adjust token embeddings for added special tokens
model.resize_token_embeddings(len(tokenizer))

# Define custom dataset
class LyricsDataset(Dataset):
    def __init__(self, encodings):
        self.encodings = encodings

    def __len__(self):
        return self.encodings.input_ids.shape[0]

    def __getitem__(self, idx):
        item = {key: val[idx] for key, val in self.encodings.items()}
        item['labels'] = item['input_ids']
        return item

# Prepare dataset
full_dataset = LyricsDataset(encodings)

# Split dataset into train and validation (90% train, 10% validation)
train_size = int(0.9 * len(full_dataset))
val_size = len(full_dataset) - train_size
train_dataset, val_dataset = random_split(full_dataset, [train_size, val_size])

# Define training arguments
training_args = TrainingArguments(
    output_dir='./results',
    overwrite_output_dir=True,
    num_train_epochs=5,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    save_steps=500,
    save_total_limit=2,
    logging_steps=100,
    learning_rate=5e-5,
    warmup_steps=200,
```

```

        prediction_loss_only=False,
        dataloader_drop_last=True,
        report_to="none"
    )

    # Initialize Trainer
    trainer = Trainer(
        model=model,
        args=training_args,
        train_dataset=train_dataset,
        eval_dataset=val_dataset,
        tokenizer=tokenizer,
    )

    # Start fine-tuning
    trainer.train()

```

<ipython-input-10-e01ad64e8908>:52: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Trainer.__init__`. Use `processing_class` instead.

```

    trainer = Trainer(
`loss_type=None` was set in the config but it is unrecognised.Using the default loss: `ForCausalLMLoss`.

```

<IPython.core.display.HTML object>

```

[ ]: TrainOutput(global_step=27000, training_loss=1.8302190602620443,
metrics={'train_runtime': 3692.5456, 'train_samples_per_second': 14.624,
'train_steps_per_second': 7.312, 'total_flos': 2.5074918752256e+16,
'train_loss': 1.8302190602620443, 'epoch': 5.0})

```

```

[ ]: # Save the fine-tuned model manually
trainer.save_model("./results")

# Save the tokenizer manually
tokenizer.save_pretrained("./results")

```

```

[ ]: ('./results/tokenizer_config.json',
'./results/special_tokens_map.json',
'./results/vocab.json',
'./results/merges.txt',
'./results/added_tokens.json')

```

2.5 Step 5: Generate Sample Lyrics Using Fine-Tuned GPT-2

```
[ ]: # Step 5: Generate sample lyrics using the fine-tuned GPT-2 model

from transformers import pipeline, GPT2Tokenizer, GPT2LMHeadModel

# Load the fine-tuned model
model_path = "./results"
model = GPT2LMHeadModel.from_pretrained(model_path)
tokenizer = GPT2Tokenizer.from_pretrained(model_path)

# Create a text generation pipeline
text_generator = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    device=0 if torch.cuda.is_available() else -1
)

# Provide a starting prompt
prompt = "Love is"

# Generate lyrics
generated = text_generator(
    prompt,
    max_length=150,
    num_return_sequences=1,
    temperature=0.7,
    top_p=0.9,
    do_sample=True,
    top_k=50,
    repetition_penalty=1.2
)

# Print the generated lyrics
print("\nGenerated Lyrics:")
print(generated[0]['generated_text'])
```

Device set to use cuda:0

Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pairs of sequences (GLUE-style) with the tokenizer you can select this strategy more precisely by providing a specific strategy to `truncation`.

Generated Lyrics:

Love is a funny thing

When love is so strong
Love can make you forget
You've got to let go of your pride and be free.

I know I'm gonna cry (I know I'll cry) Oh why won't you forgive me
For loving someone like that?(For loving someone like this)... Kentucky! A B C D
E F G H R O U N T S L Y X - Love is such an easy game, baby yeah
Yeah

It's not what it seems
'Cause love is all about trust
And the only time that I'm gonna let you go
Is when you're hurting too much or being untrue
Don't put me on that high
Just leave

2.6 Step 6.1: Generate Multiple Lyrics Samples

```
[ ]: # Generate 3 different lyrics samples using different prompts
```

```
prompts = [  
    "Broken dreams",  
    "Summer nights",  
    "Dancing under the rain"  
]  
  
for prompt in prompts:  
    generated = text_generator(  
        prompt,  
        max_length=150,  
        num_return_sequences=1,  
        temperature=0.7,  
        top_p=0.9,  
        do_sample=True,  
        top_k=50,  
        repetition_penalty=1.2  
    )  
    print(f"\nPrompt: {prompt}")  
    print("Generated Lyrics:")  
    print(generated[0]['generated_text'])  
    print("-" * 100)
```

Prompt: Broken dreams
Generated Lyrics:
Broken dreams, broken hearts
Can never heal the scars of a lifetime torn apart

I'm alone and I don't know why
I feel like crying all the time
And every night it's raining tears on my pillow Banks B - O'Hara (B-0)

Broke dreams, broke heartaches
Can never mend the hurt that is left in our love
I'm alone but I don't know why
I feel like dying each day
And every night it rains tears on my bed... Broke Dreams! Broken Heartache.
etc..etc

(x2) The sky is black and gray
My life has just begun to fade away now so lonely inside yeah
There is nothing else

Prompt: Summer nights

Generated Lyrics:

Summer nights are calling, and the leaves begin to fall
I close my eyes as I dream of you
And then I'm alone again
As if in a tunnel deep inside me
My thoughts go on forever
They travel through time

My love is all around me, my heart's with her every step.
When I think about our wedding day
I know that I should be there for her too. Seraphim - Piano Concerto No 1
(Pianoforte)

... And I remember how she held me tight each night when we danced so sweetly
away at home over yonder hill...(end)...} My love is all round me, my heart's
where she keeps me every hour or two

Prompt: Dancing under the rain

Generated Lyrics:

Dancing under the rain
In my car
I'm alone again
I know that I'll always be
The only one for you

Darling, I love your eyes
And all of those things about me
That make me want to hold you so tight. (Hold you so tightly) Yeah!

(Lyrics: brecht) -- Instrumental and Video Sized... Lyrics by brecht Music :
Simple Minds Reproduced Without Permission.(C) EMI Publishing Ltd... Right up
front there's a message on our screen from Simple mind s' dot com where we can
get in touch with you if you'd like us too but please don't send us any photos
because they might get

```
trainer.save_model("./results") tokenizer.save_pretrained("./results")
```

Conclusion

In this project, we fine-tuned the GPT-2 Medium model on an English lyrics dataset. The model was trained for 5 epochs and achieved a final training loss of approximately 1.83. After fine-tuning, the model was able to generate creative and meaningful English song lyrics based on different prompts. The project demonstrates the capability of transformer-based language models to adapt to domain-specific text generation tasks with minimal fine-tuning.