Names:

Paige Hodgkinson
Sid Padmanabhuni
Chun-Wei Pan

San Francisco State University

Course CSC 671 Deep Learning
Team Project: Team 1
LSTM for Text Generation

Introduction

Natural language generation is a crucial area of deep learning, aiming to create human-quality text based on the style and content of the data on which we train. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, are typically used for this. However, LSTMs face limitations in capturing long-range dependencies within text. To address this, bidirectional LSTMs and attention mechanisms have emerged as the most effective techniques, and we utilized both in the design of this model.

Our team here details the development and evaluation of a text generation LSTM model. The model was trained on a subset of the WikiText-103 dataset and assessed for its ability to generate coherent and contextually relevant text.

Methodology and Design

To prepare the data, a subset of the WikiText-103 dataset was selected and preprocessed. This involved filtering out unnecessary symbols, tokenizing the text into words, and creating sequences of fixed length. Each sequence served as input to the model, with the next word in the sequence as the target output.

```
# Import libraries

import re

from tensorflow.keras.preprocessing.text import Tokenizer



# Load a subset of WikiText-103 dataset

dataset = load_dataset("wikitext", "wikitext-103-raw-v1",
split="train[:15%]")  # Reduced dataset size
```

```
# Filter out unwanted symbols

text = " ".join(dataset['text'])

filtered_text = re.sub(r'[^A-Za-z0-9.,\'\s-]', '', text)

print(f"Filtered text length: {len(filtered_text)} characters")


# Tokenize text

tokenizer = Tokenizer()  # Word-level tokenizer

tokenizer.fit_on_texts([filtered_text])

total_words = len(tokenizer.word_index) + 1

print(f"Total unique words: {total_words}")
```

The primary function of the model is a bidirectional LSTM network that processes the input sequence in both forward and backward directions at once and then combines the results for more accuracy, capturing long-range dependencies. An attention mechanism is employed to weigh the importance of different parts of the input sequence, enabling the model to focus on relevant word associations. The final output layer generates the probability distribution over the vocabulary for the next word.

The model was trained using the cross-entropy loss function and the Adam optimizer. To prevent overfitting and improve training stability, gradient clipping and learning rate scheduling were implemented. The model was trained on a reduced dataset with a batch size of 512 and a limited number of epochs. Gradient accumulation was also added to improve training efficiency due to slow training.

```
# Define the LSTM Model

class LSTMTextGenerationModel(nn.Module):

  def __init__(self, vocab_size, embedding_dim=150, hidden_dim=512, output_dim=None):

    super(LSTMTextGenerationModel, self).__init__()

    self.embedding = nn.Embedding(vocab_size, embedding_dim)
    self.lstm = nn.LSTM(embedding_dim, hidden_dim, num_layers=2, batch_first=True, dropout=0.2, bidirectional=True)

    self.attention_fc = nn.Linear(hidden_dim * 2, 1)

    self.fc = nn.Linear(hidden_dim * 2, output_dim)  # Account for bidirectionality


  def attention(self, lstm_out):

    attention_weights = torch.tanh(self.attention_fc(lstm_out))

    attention_weights = F.softmax(attention_weights, dim=1)

    weighted_output = torch.sum(attention_weights * lstm_out, dim=1)

    return weighted_output


  def forward(self, x):

    x = self.embedding(x)

    x, _ = self.lstm(x)

    attention_out = self.attention(x)

    x = self.fc(attention_out)

    return x
```

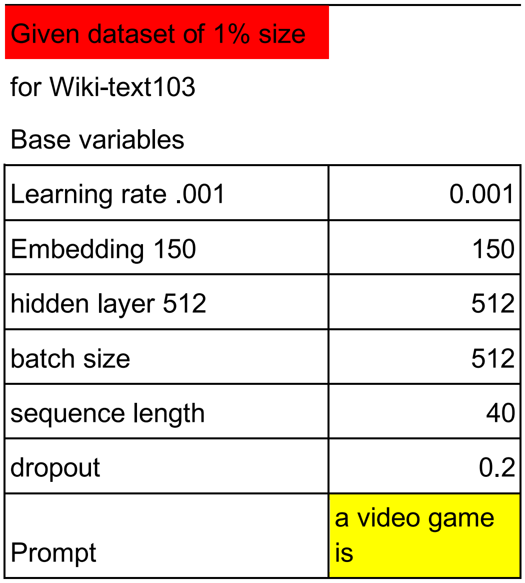
To generate text, the model takes a user-provided seed text as input and iteratively predicts the next word. The output *logits* are scaled by a *temperature* parameter to control the randomness of the generated text, and top-k sampling is used to select the most probable words (also for efficiency). This process is repeated until a given number of words is generated.

```
# Define the text generation function with temperature scaling and top-k sampling

def generate_text(seed_text, next_words=100, temperature=0.7, top_k=10):

  # Filter non-essential characters

  seed_text = re.sub(r'[^A-Za-z0-9.,\'\s-]', '', seed_text)
```

```
generated_text =
```

The model's performance was evaluated based on its ability to generate coherent and contextually relevant text. While we utilized training loss and perplexity metric to assess the model's learning progress, human evaluation at such a small scale of development was important. Visually assessing the coherence of the text generated results was our primary indicator of success (forgoing a BLEU metric altogether).

Results



Given dataset of 1% size

for Wiki-text103

Base variables

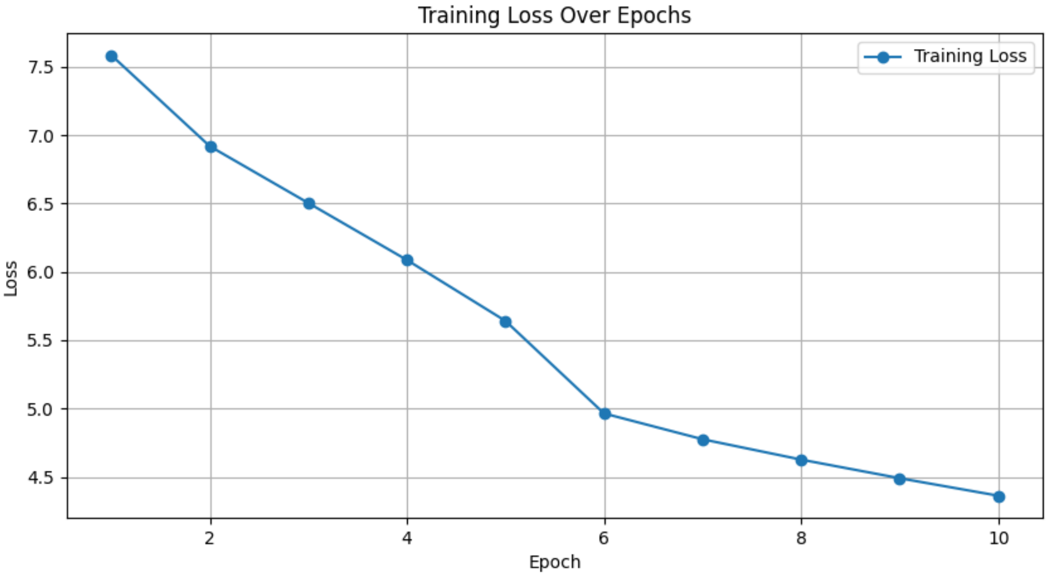
Learning rate .001	0.001
Embedding 150	150
hidden layer 512	512
batch size	512
sequence length	40
dropout	0.2
Prompt	a video game is

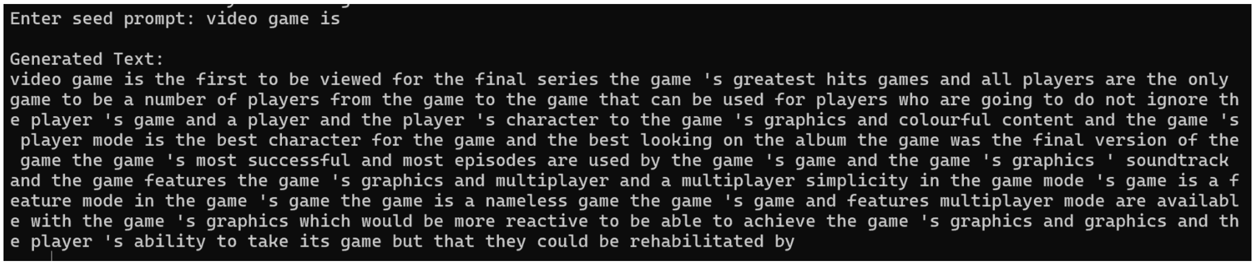
Hyper Parameter Testing Grid	All training on: NVIDIA RTX 4090

Hyperparameter	Values	Time (ns/ms?)	Starting Loss @ (Epoch 1/10)	Final Loss @ (Epoch 10/10)	Perplexity Score	Output Figure Name
Initial Learning Rate	0.001	8m, 42s	7.2467	3.1587	1.01	"Base"
	0.0005	8m, 44s	7.5849	4.3629	1.01	"LR0.0005"
	0.0001	8m, 42s	7.8253	6.4655	1.01	"LR0.0001"
Embedding Dimension	100	8m, 33s	7.4868	3.19	1.01	"ED100"
	150	8m, 42s	7.2467	3.1587	1.01	"Base"
	200	8m, 44s	7.3953	2.805	1.01	"ED200"
Hidden Dimension	128	2m, 47s	7.6748	5.6147	1.01	"HD128"
	256	4m, 46s	7.5804	4.6137	1.01	"HD256"
	512	8m, 42s	7.2467	3.1587	1.01	"Base"
Dropout Rate	0.1	8m, 44s	7.4151	2.5499	1	"DR0.1"
	0.2	8m, 42s	7.2467	3.1587	1.01	"Base"
	0.3	8m, 44s	7.4752	3.2731	1.01	"DR0.3"
Sequence Length	20	5m, 41s	7.4171	2.8938	1.01	"SL20"
	40	8m, 42s	7.2467	3.1587	1.01	"Base"
	80	14m, 38s	7.4542	2.9955	1.01	"SL80"
Batch Size	64	26m, 17s	7.1781	2.3529	1.04	"BS64"
	256	11m, 12s	7.348	2.6544	1.01	"BS256"
	512	8m, 42s	7.2467	3.1587	1.01	"Base"

Table of Hyperparameter Search Results

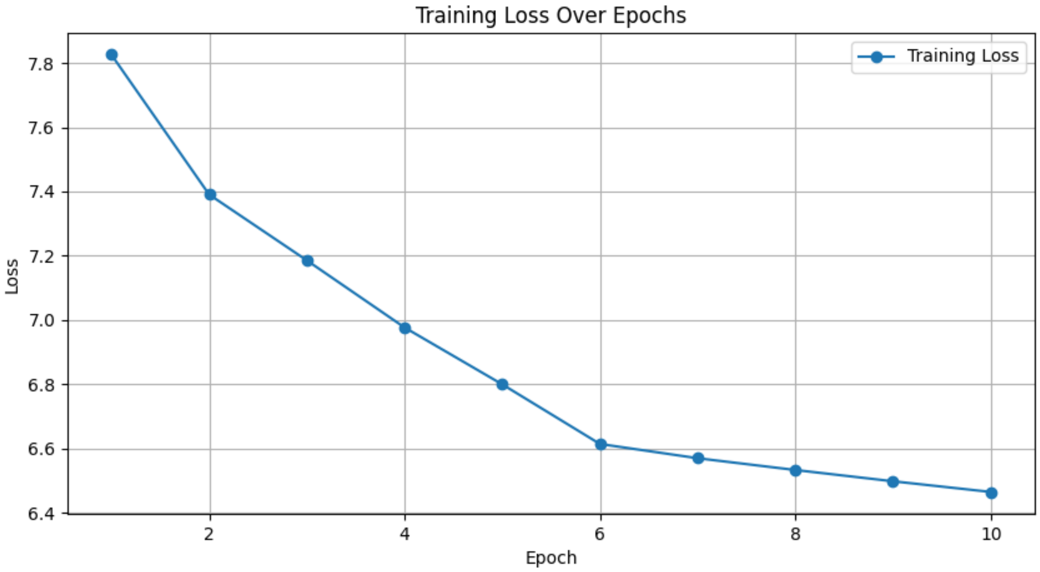
LR0.0005

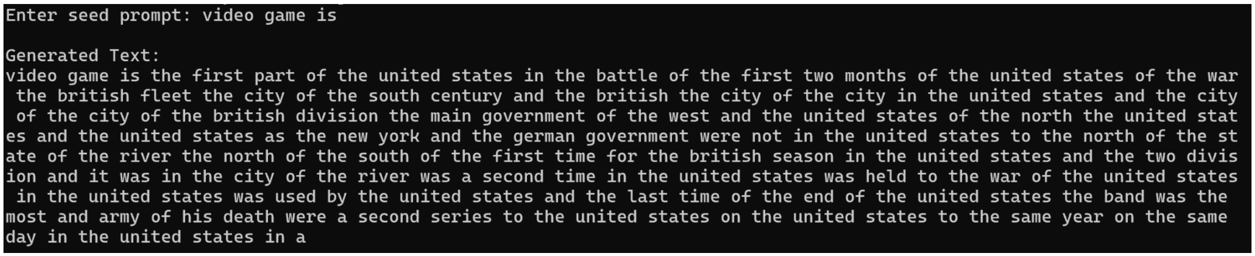




```
Enter seed prompt: video game is

Generated Text:
video game is the first to be viewed for the final series the game 's greatest hits games and all players are the only
game to be a number of players from the game to the game that can be used for players who are going to do not ignore th
e player 's game and a player and the player 's character to the game 's graphics and colourful content and the game 's
 player mode is the best character for the game and the best looking on the album the game was the final version of the
 game the game 's most successful and most episodes are used by the game 's game and the game 's graphics ' soundtrack
and the game features the game 's graphics and multiplayer and a multiplayer simplicity in the game mode 's game is a f
eature mode in the game 's game the game is a nameless game the game 's game and features multiplayer mode are availabl
e with the game 's graphics which would be more reactive to be able to achieve the game 's graphics and graphics and th
e player 's ability to take its game but that they could be rehabilitated by
```

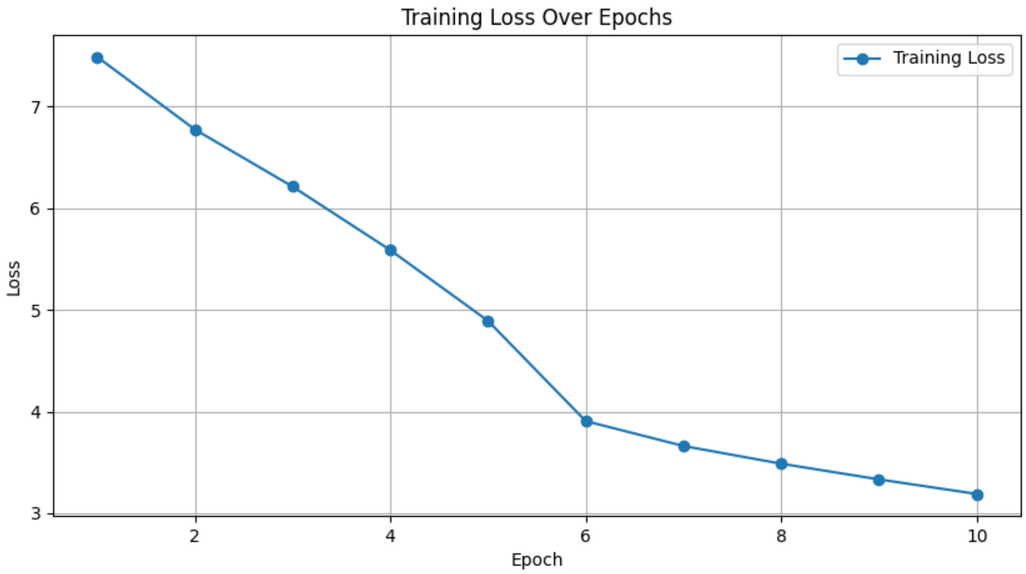
LR0.0001

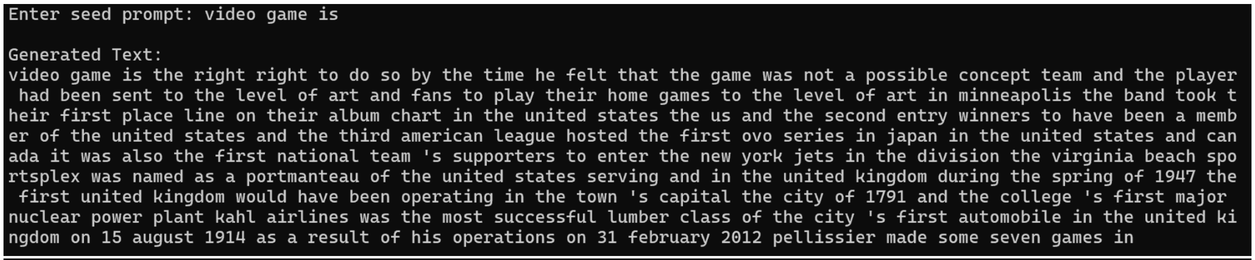




```
Enter seed prompt: video game is

Generated Text:
video game is the first part of the united states in the battle of the first two months of the united states of the war
 the british fleet the city of the south century and the british the city of the city in the united states and the city
 of the city of the british division the main government of the west and the united states of the north the united stat
es and the united states as the new york and the german government were not in the united states to the north of the st
ate of the river the north of the south of the first time for the british season in the united states and the two divis
ion and it was in the city of the river was a second time in the united states was held to the war of the united states
 in the united states was used by the united states and the last time of the end of the united states the band was the
most and army of his death were a second series to the united states on the united states to the same year on the same
day in the united states in a
```

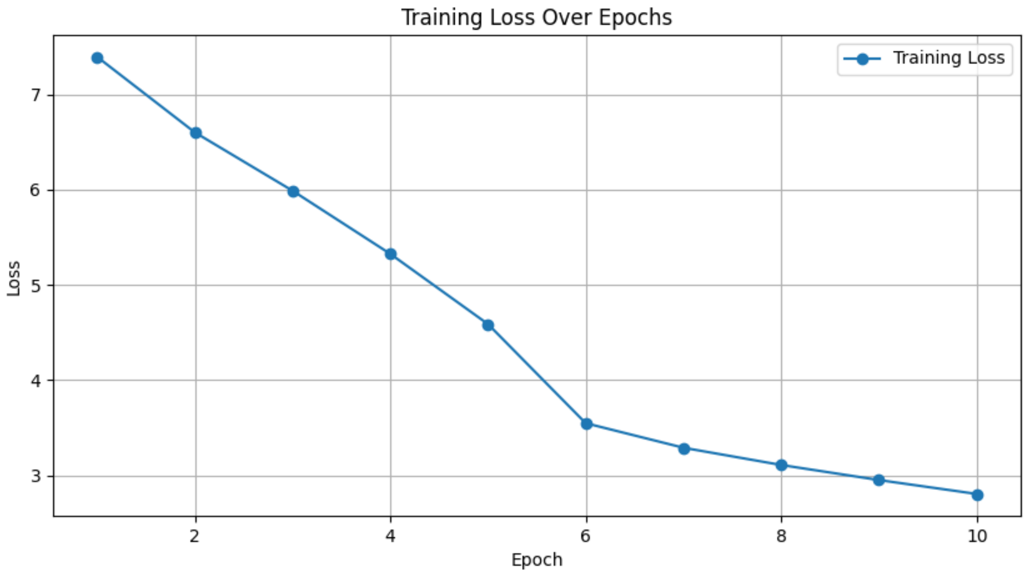
ED100

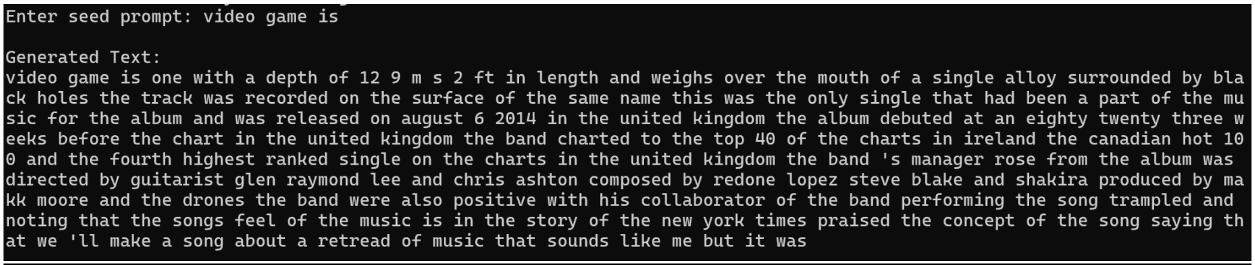




```
Enter seed prompt: video game is

Generated Text:
video game is the right right to do so by the time he felt that the game was not a possible concept team and the player
 had been sent to the level of art and fans to play their home games to the level of art in minneapolis the band took t
heir first place line on their album chart in the united states the us and the second entry winners to have been a memb
er of the united states and the third american league hosted the first ovo series in japan in the united states and can
ada it was also the first national team 's supporters to enter the new york jets in the division the virginia beach spo
rtsplex was named as a portmanteau of the united states serving and in the united kingdom during the spring of 1947 the
 first united kingdom would have been operating in the town 's capital the city of 1791 and the college 's first major
nuclear power plant kahl airlines was the most successful lumber class of the city 's first automobile in the united ki
ngdom on 15 august 1914 as a result of his operations on 31 february 2012 pellissier made some seven games in
```

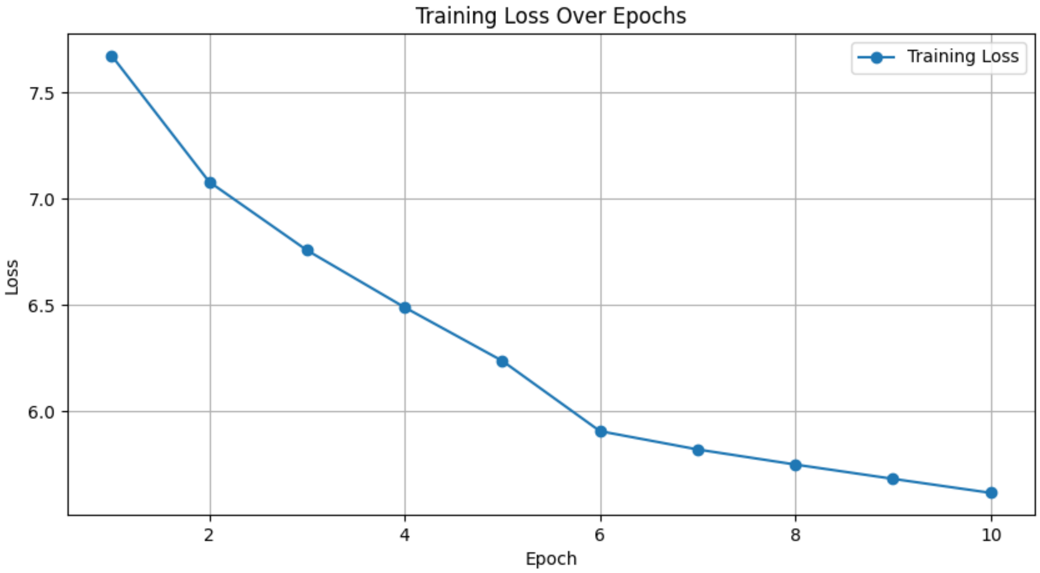
ED200

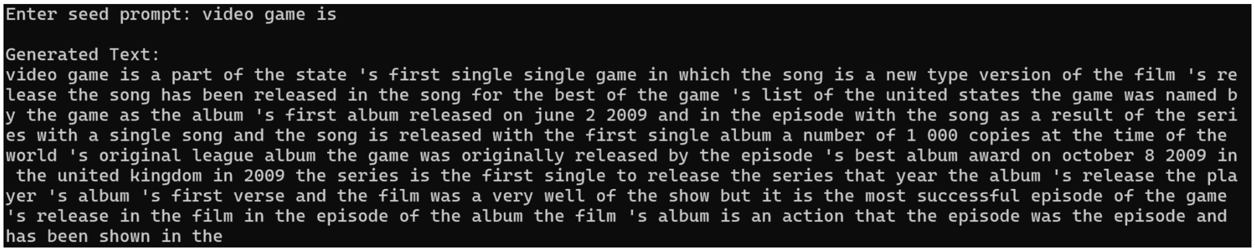




```
Enter seed prompt: video game is

Generated Text:
video game is one with a depth of 12 9 m s 2 ft in length and weighs over the mouth of a single alloy surrounded by bla
ck holes the track was recorded on the surface of the same name this was the only single that had been a part of the mu
sic for the album and was released on august 6 2014 in the united kingdom the album debuted at an eighty twenty three w
eeks before the chart in the united kingdom the band charted to the top 40 of the charts in ireland the canadian hot 10
0 and the fourth highest ranked single on the charts in the united kingdom the band 's manager rose from the album was
directed by guitarist glen raymond lee and chris ashton composed by redone lopez steve blake and shakira produced by ma
kk moore and the drones the band were also positive with his collaborator of the band performing the song trampled and
noting that the songs feel of the music is in the story of the new york times praised the concept of the song saying th
at we 'll make a song about a retread of music that sounds like me but it was
```

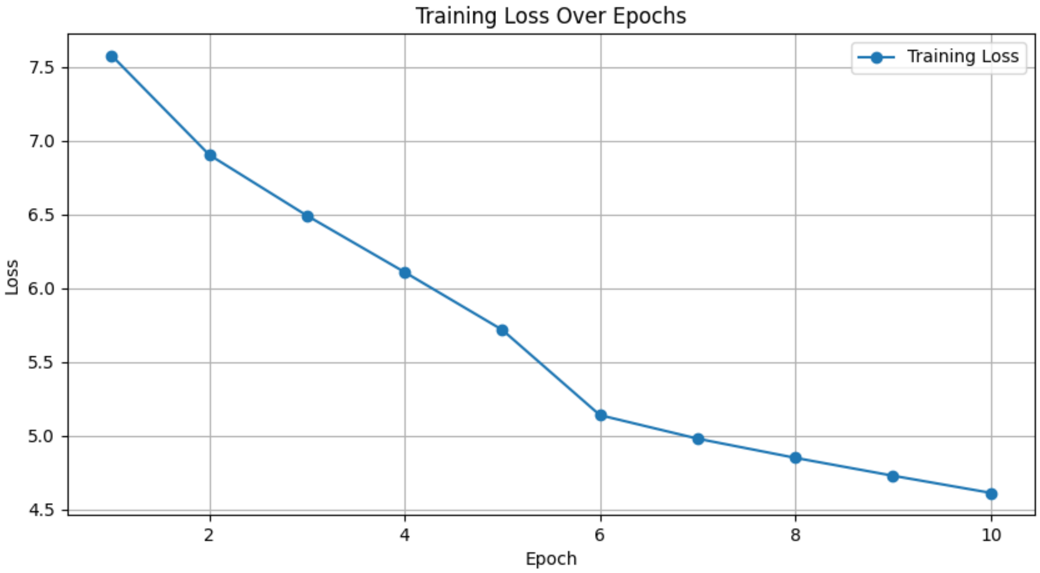
HD128

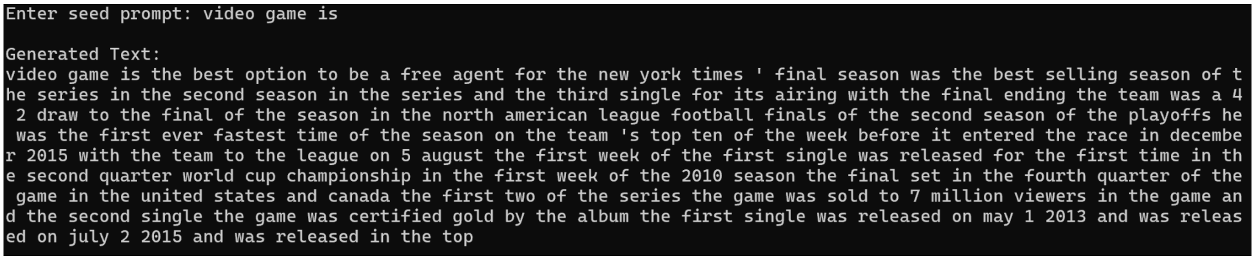




```
Enter seed prompt: video game is

Generated Text:
video game is a part of the state 's first single single game in which the song is a new type version of the film 's re
lease the song has been released in the song for the best of the game 's list of the united states the game was named b
y the game as the album 's first album released on june 2 2009 and in the episode with the song as a result of the seri
es with a single song and the song is released with the first single album a number of 1 000 copies at the time of the
world 's original league album the game was originally released by the episode 's best album award on october 8 2009 in
 the united kingdom in 2009 the series is the first single to release the series that year the album 's release the pla
yer 's album 's first verse and the film was a very well of the show but it is the most successful episode of the game
's release in the film in the episode of the album the film 's album is an action that the episode was the episode and
has been shown in the
```

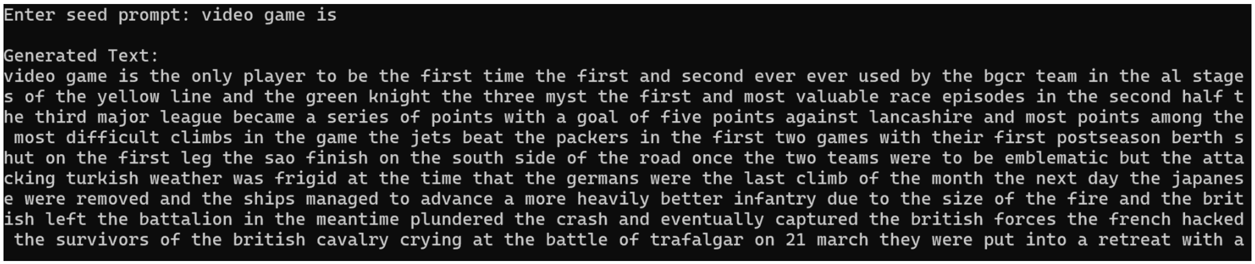
HD256



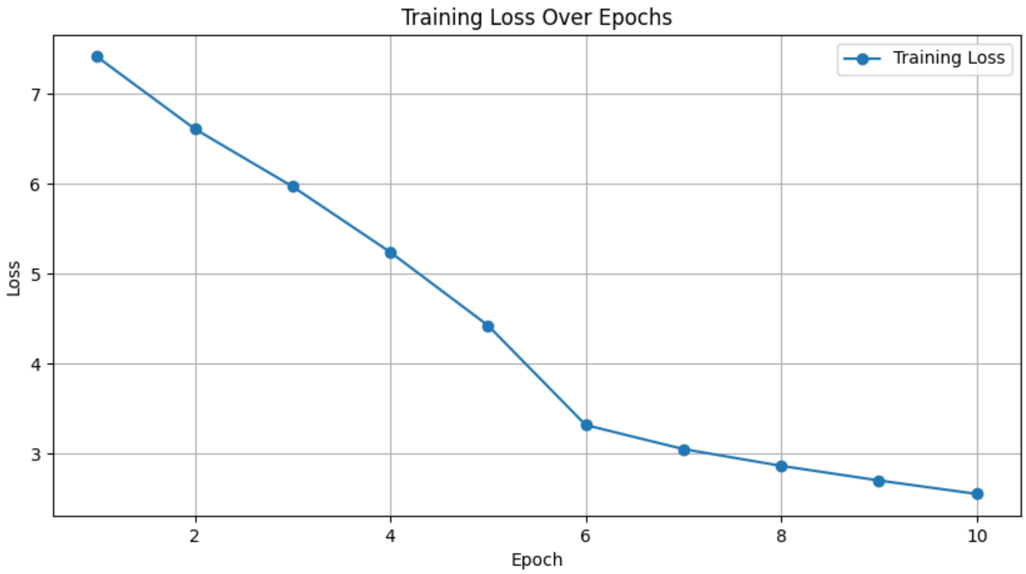


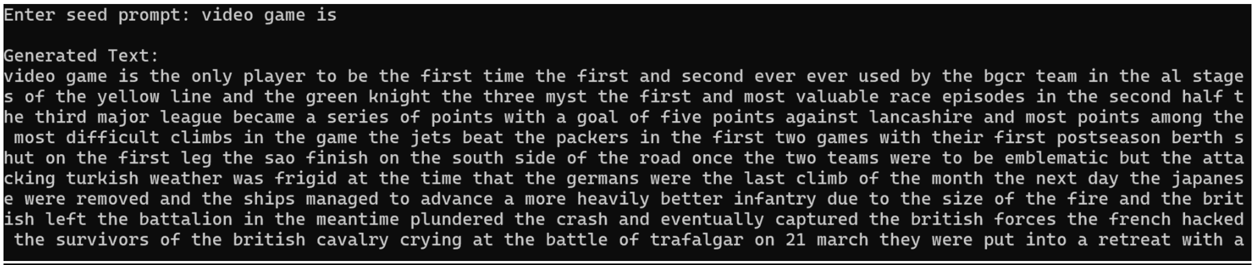
```
Enter seed prompt: video game is

Generated Text:
video game is the best option to be a free agent for the new york times ' final season was the best selling season of t
he series in the second season in the series and the third single for its airing with the final ending the team was a 4
 2 draw to the final of the season in the north american league football finals of the second season of the playoffs he
 was the first ever fastest time of the season on the team 's top ten of the week before it entered the race in decembe
r 2015 with the team to the league on 5 august the first week of the first single was released for the first time in th
e second quarter world cup championship in the first week of the 2010 season the final set in the fourth quarter of the
 game in the united states and canada the first two of the series the game was sold to 7 million viewers in the game an
d the second single the game was certified gold by the album the first single was released on may 1 2013 and was releas
ed on july 2 2015 and was released in the top
```

DR0.1



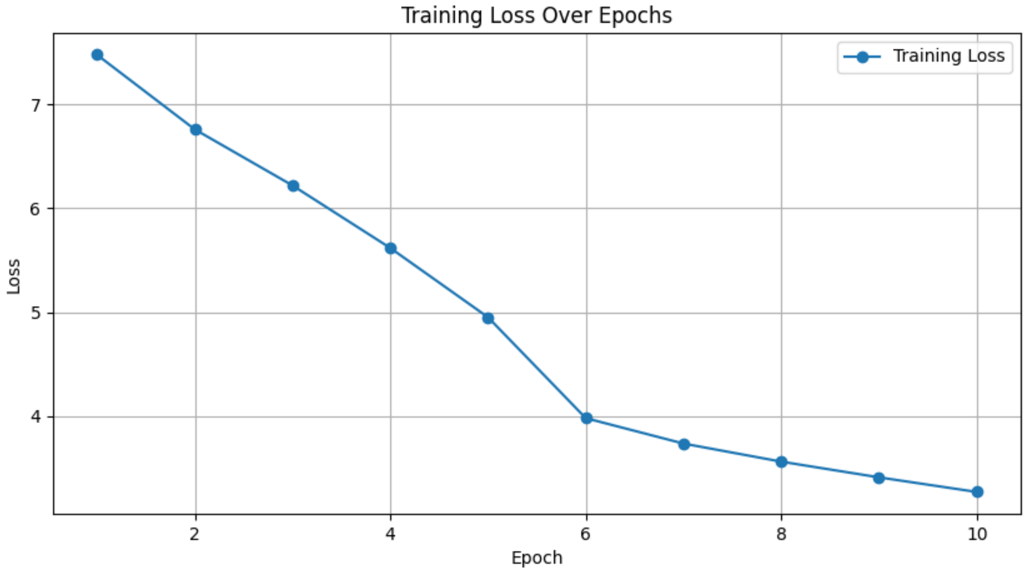
```
Enter seed prompt: video game is

Generated Text:
video game is the only player to be the first time the first and second ever ever used by the bgcr team in the al stage
s of the yellow line and the green knight the three myst the first and most valuable race episodes in the second half t
he third major league became a series of points with a goal of five points against lancashire and most points among the
 most difficult climbs in the game the jets beat the packers in the first two games with their first postseason berth s
hut on the first leg the sao finish on the south side of the road once the two teams were to be emblematic but the atta
cking turkish weather was frigid at the time that the germans were the last climb of the month the next day the japanes
e were removed and the ships managed to advance a more heavily better infantry due to the size of the fire and the brit
ish left the battalion in the meantime plundered the crash and eventually captured the british forces the french hacked
 the survivors of the british cavalry crying at the battle of trafalgar on 21 march they were put into a retreat with a
```

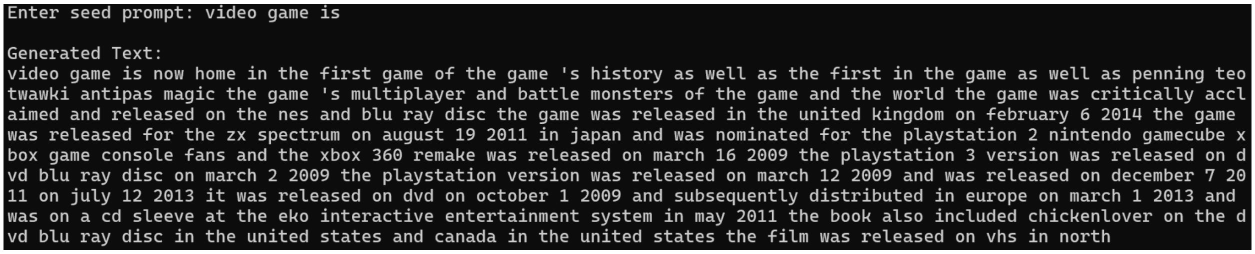




```
Enter seed prompt: video game is

Generated Text:
video game is the only player to be the first time the first and second ever ever used by the bgcr team in the al stage
s of the yellow line and the green knight the three myst the first and most valuable race episodes in the second half t
he third major league became a series of points with a goal of five points against lancashire and most points among the
 most difficult climbs in the game the jets beat the packers in the first two games with their first postseason berth s
hut on the first leg the sao finish on the south side of the road once the two teams were to be emblematic but the atta
cking turkish weather was frigid at the time that the germans were the last climb of the month the next day the japanes
e were removed and the ships managed to advance a more heavily better infantry due to the size of the fire and the brit
ish left the battalion in the meantime plundered the crash and eventually captured the british forces the french hacked
 the survivors of the british cavalry crying at the battle of trafalgar on 21 march they were put into a retreat with a
```

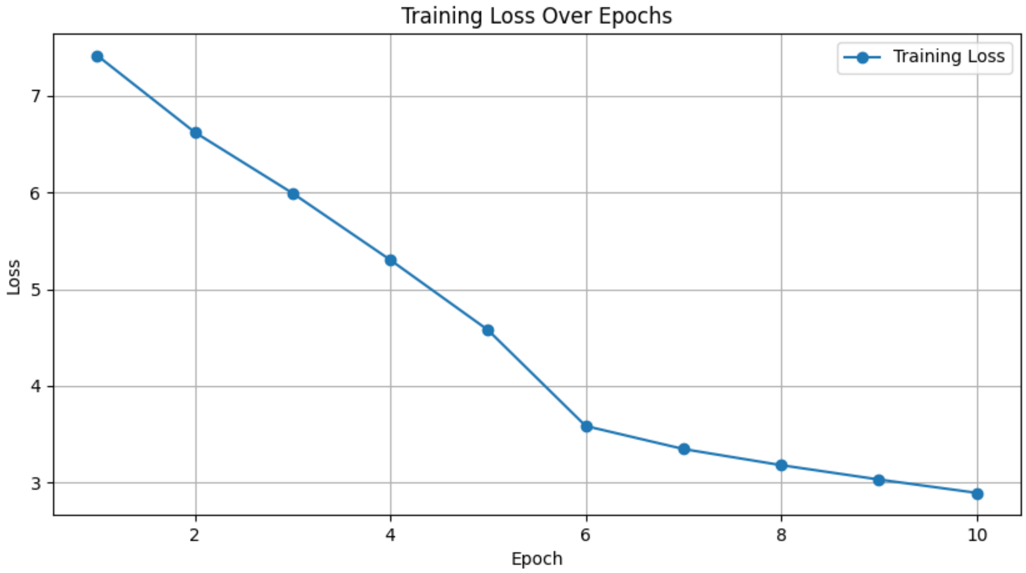
DR0.3

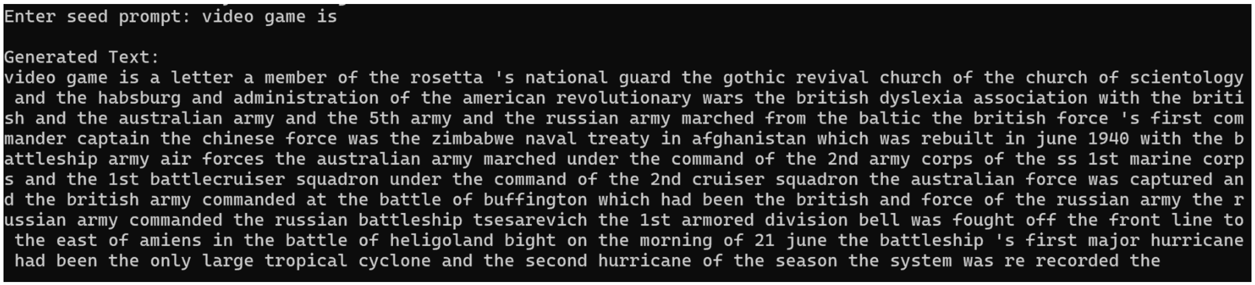




```
Enter seed prompt: video game is

Generated Text:
video game is now home in the first game of the game 's history as well as the first in the game as well as penning teo
twawki antipas magic the game 's multiplayer and battle monsters of the game and the world the game was critically accl
aimed and released on the nes and blu ray disc the game was released in the united kingdom on february 6 2014 the game
was released for the zx spectrum on august 19 2011 in japan and was nominated for the playstation 2 nintendo gamecube x
box game console fans and the xbox 360 remake was released on march 16 2009 the playstation 3 version was released on d
vd blu ray disc on march 2 2009 the playstation version was released on march 12 2009 and was released on december 7 20
11 on july 12 2013 it was released on dvd on october 1 2009 and subsequently distributed in europe on march 1 2013 and
was on a cd sleeve at the eko interactive entertainment system in may 2011 the book also included chickenlover on the d
vd blu ray disc in the united states and canada in the united states the film was released on vhs in north
```

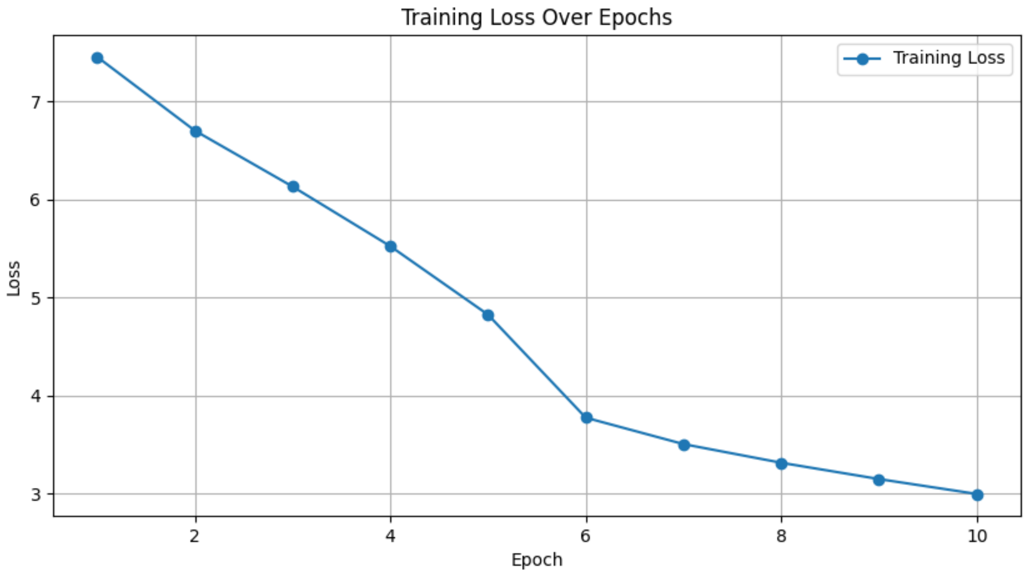
SL20

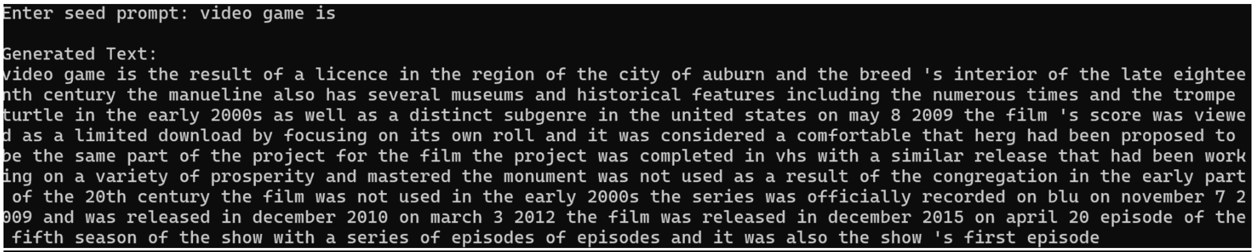




```
Enter seed prompt: video game is

Generated Text:
video game is a letter a member of the rosetta 's national guard the gothic revival church of the church of scientology
 and the habsburg and administration of the american revolutionary wars the british dyslexia association with the briti
sh and the australian army and the 5th army and the russian army marched from the baltic the british force 's first com
mander captain the chinese force was the zimbabwe naval treaty in afghanistan which was rebuilt in june 1940 with the b
attleship army air forces the australian army marched under the command of the 2nd army corps of the ss 1st marine corp
s and the 1st battlecruiser squadron under the command of the 2nd cruiser squadron the australian force was captured an
d the british army commanded at the battle of buffington which had been the british and force of the russian army the r
ussian army commanded the russian battleship tsesarevich the 1st armored division bell was fought off the front line to
 the east of amiens in the battle of heligoland bight on the morning of 21 june the battleship 's first major hurricane
 had been the only large tropical cyclone and the second hurricane of the season the system was re recorded the
```

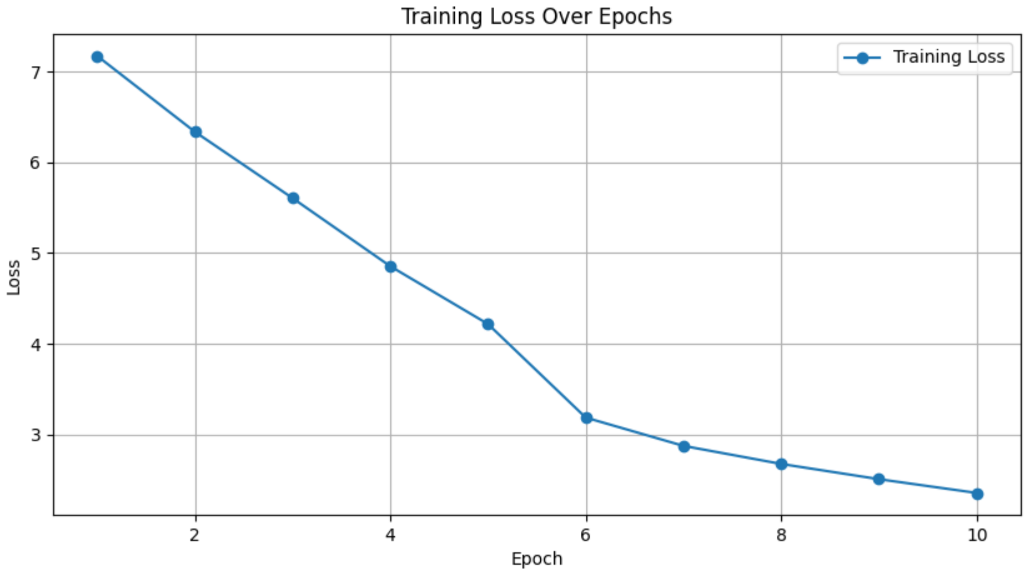
SL80



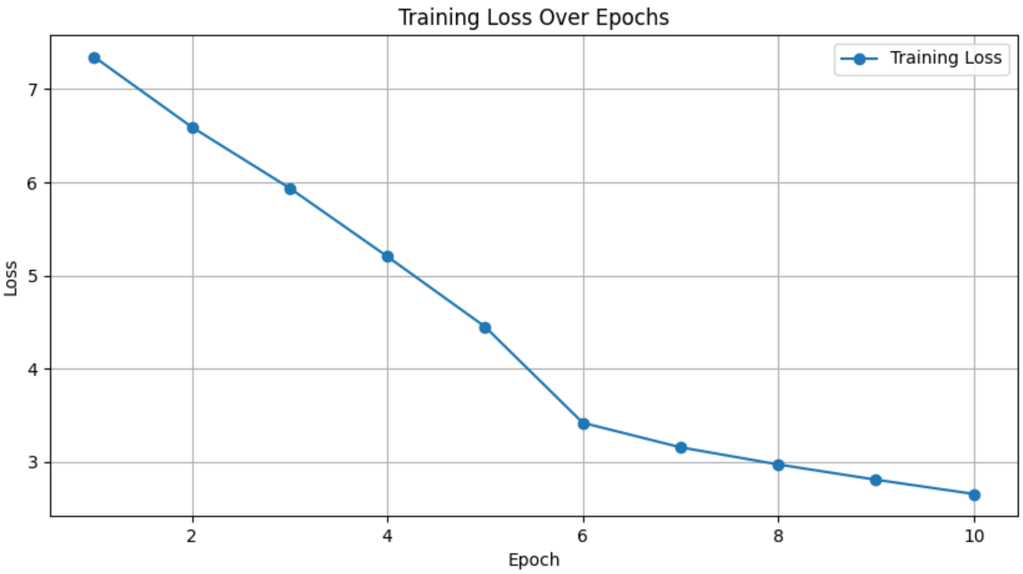


```
Enter seed prompt: video game is

Generated Text:
video game is the result of a licence in the region of the city of auburn and the breed 's interior of the late eightee
nth century the manueline also has several museums and historical features including the numerous times and the trompe
turtle in the early 2000s as well as a distinct subgenre in the united states on may 8 2009 the film 's score was viewe
d as a limited download by focusing on its own roll and it was considered a comfortable that herg had been proposed to
be the same part of the project for the film the project was completed in vhs with a similar release that had been work
ing on a variety of prosperity and mastered the monument was not used as a result of the congregation in the early part
 of the 20th century the film was not used in the early 2000s the series was officially recorded on blu on november 7 2
009 and was released in december 2010 on march 3 2012 the film was released in december 2015 on april 20 episode of the
 fifth season of the show with a series of episodes of episodes and it was also the show 's first episode
```

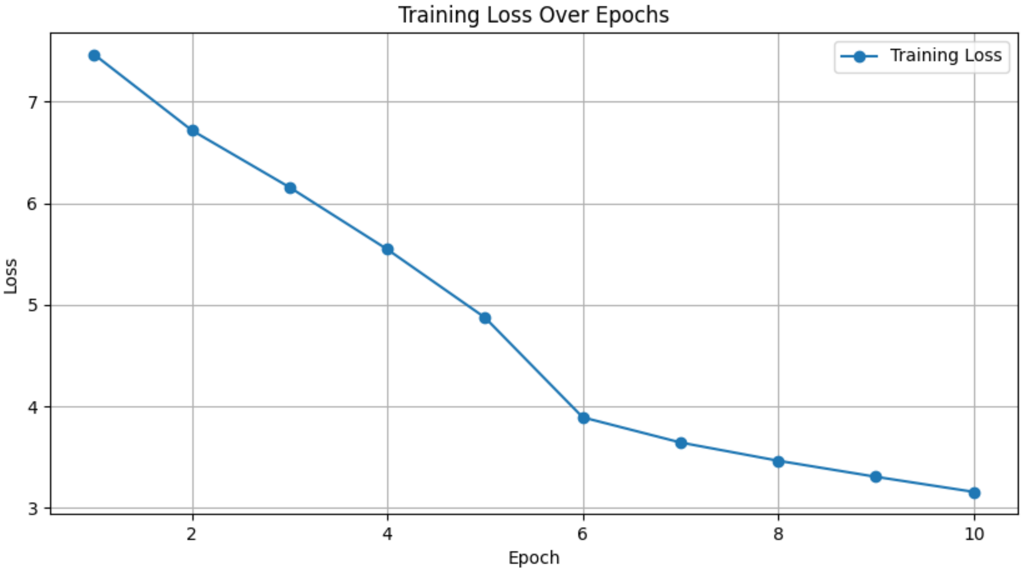
BS64



BS256



Base



Hyperparameters: Initial Learning Rate .001, Embedding Dimension 150, Hidden Dimension 512, Dropout Rate .2, Sequence Length 40, Batch Size 512

Training on 1% of dataset, hyperparameter grid search results.

A training with the initial set of hyperparameters at 30% of the dataset.

Sample Training Result:

Enter seed prompt: A video game is

Generated Text:

A video game is released in japan on march 2 2017 in japan the playstation vita playstation network windows playstation portable psp version is available in the united kingdom for microsoft windows playstation 3 playstation 3 and xbox 360 versions in the same version as the game 's first vita version of the game the ps2 version is based on two discs the game 's title version was released in japan on september 26 2003 but was released on september 30 2006 in europe and north america in north america on april 24 2005 a japanese version of the game was released in north america on december 4 2006 and november 24 2006 the north american version of the game was ported to the playstation 2 version of the game boy advance the game 's music was released in japan on may 25 2006 on june 22 2008 in europe and on july 6 2007 the game is the last game in the series to feature a demo and play for the player character development system which includes a new story and story mode that allows the player to complete a game the game was created and developed by crytek insomniac and published by

Discussion of Results

Perplexity score was an ineffective indicator because the scores for all hyperparameters were all almost exactly the same. (ranging from 1 to 1.04, usually 1.01). Final Loss showed much more notable indications between results. Increasing the dimensions of the embedding and hidden layers (and therefore more connections) caused more effective convergence (and therefore accuracy). Similarly, reducing other hyperparameters also caused less loss and more accuracy, but required more training time. Because massive computing power was necessary for running optimization within reasonable time, for our 30% training we only had 2 viable computers for training on. For 18 parameters, 18x18 (324 runs) gridspace would be necessary to capture the complex effects of all variables on each other, which would need approximately 162 hours of runtime.