1 Problem 2

1.1 Objective

The objective of part 2 of this assignment is to implement a Recurrent Neural Network (RNN) to create an English language character generator capable of building semi-coherent English sentences from scratch, by building them up **character-by-character**.

For training the model, we will use the complete version of J. K. Rowling's book "Harry Potter and the Chamber of Secrets".

A deep learning training model can be trained to generate text automatically, character-by-character, by showing the model many training examples so that it can learn a pattern between text inputs and potential character outputs. With this type of text generation, each input is a string of valid characters like this one:

"THE QUICK BROW"

The corresponding output is the next character in the sentence is 'n'. We need to show a model many such examples for it to make reasonable predictions.

1.2 Data Loading & Processing

The file is opened and loaded using encoding='utf-8' to ensure that the text is correctly interpreted, especially when it contains non-ASCII characters. This is important because:

- **Unicode Standard**: UTF-8 is a widely used character encoding that supports all Unicode characters. This includes letters, symbols, and emojis from nearly all writing systems.
- **Avoiding Errors**: Without specifying utf-8, Python might use the default system encoding, which can lead to errors when reading files with diverse character sets. This is because different systems might have different default encodings (like Windows often uses CP1252).
- **Consistency**: By explicitly using UTF-8, this ensures my script behaves consistently across different systems and environments. This is crucial for portability and avoiding bugs related to character encoding.
- **Non-ASCII characters**: Books like *Harry Potter and the Chamber of Secrets* contain various punctuation marks, foreign words, and special characters that are outside the basic ASCII range. Using UTF-8 ensures all these characters are correctly read and processed.

After the file is opened and read, all the characters are converted to lowercase for consistency. The total number of characters in the original text is: **531708** characters.

Next, unnecessary characters that do not contribute to the understanding of the text can be removed to clean and preprocess the data. These include:

- Whitespace: Extra spaces and newlines
- Numbers: In some cases, numbers might be considered unnecessary unless they are critical to the context.
- Non-ASCII Characters: Especially in cases where the text is expected to be in a particular language or script.
- Punctuation Marks: Unless needed for sentiment or context.

- Stop word removal is not carried out for character generation because the model is learning the relationship between characters.
    - It needs to learn the statistical patterns of how letters follow each other, including the spaces between words.
    - Stop words are vital for forming grammatical structures and sentence flow.
    - The model needs to learn proper sentence structure and grammar, and stop words are a crucial part of it.

I used the regex package to keep lowercase a-z characters and specified punctuation like periods, commas, exclamation, question and quotation marks, apostrophes, colons and parenthesis.

Once the text has been cleaned, we find the list of unique characters and punctuations in the sanitized text. The cleansed text now has a total of **502869** characters, with **37** unique characters.

To prepare the data into training text and labels (X & y), the "sliding window" method is used, with a step size of 3. Too small a step size e.g. 1, generates many overlapping sequences. Setting it to 'seq-length', which is the size of your input sequence, leads to no overlap, will result in completely distinct sequences, which is something we also do not want.

Data preparation is followed by one hot encoding. First, we create a dictionary mapping the characters to indices and vice versa. The training text (X) and labels (y) are passed to the one hot encoding function. To illustrate this better, say we have a small dataset with the following input sequences and next characters:

- **Input Sequences (X)**: ["abc", "bcd", "cde"]
- **Next Characters (y)**: ["d", "e", "f"]

Assume the unique characters in our text are:

- **Unique Characters**: ['a', 'b', 'c', 'd', 'e', 'f']

We'll use these mappings:

- **Character to Index**:
    - char_to_index = {'a': 0, 'b': 1, 'c': 2, 'd': 3, 'e': 4, 'f': 5}

The expected output after passing through the one hot encoding function will be:

X_encoded

The X_encoded array will be a 3D array with one-hot encoded sequences:

```
[[[1 0 0 0 0 0]  # "a" -> [1 0 0 0 0 0]
  [0 1 0 0 0 0]  # "b" -> [0 1 0 0 0 0]
  [0 0 1 0 0 0]] # "c" -> [0 0 1 0 0 0]

 [[0 1 0 0 0 0]  # "b" -> [0 1 0 0 0 0]
  [0 0 1 0 0 0]  # "c" -> [0 0 1 0 0 0]
  [0 0 0 1 0 0]] # "d" -> [0 0 0 1 0 0]

 [[0 0 1 0 0 0]  # "c" -> [0 0 1 0 0 0]
  [0 0 0 1 0 0]  # "d" -> [0 0 0 1 0 0]
  [0 0 0 0 1 0]]]# "e" -> [0 0 0 0 1 0]
```

y_encoded

The y_encoded array will be a 2D array with one-hot encoded next characters:

```
[[0 0 0 1 0 0]  # "d" -> [0 0 0 1 0 0]
 [0 0 0 0 1 0]  # "e" -> [0 0 0 0 1 0]
```

```
[0 0 0 0 0 1]] # "f" -> [0 0 0 0 0 1]
```

In each row of X_encoded, only one element in each column is 1, representing the specific character. Similarly, y_encoded has a single element set to 1 for each next character.

This X_encoded and y_encoded format is what the model uses for training, allowing it to learn patterns from these one-hot encoded sequences.

The X_encoded and y_encoded data is then split into the X_train, X_test, y_train and y_test using the **train_test_split()** from sklearn. We specify a test size of 0.2 , that means 20% of the dataset was used for testing, leaving 80% for training

1.3 Sequence Generator Models

Two models were built for this exercise.

The first model consists of 2 LSTM layers followed by a dense layer. The first LSTM layer uses **128** units, the second **64** units with a final dense layer having **37** units (which is the number of unique characters from our cleaned dataset). A summary of the model is shown in Figure 1.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_2 (LSTM)               (None, 50, 128)           84992

 lstm_3 (LSTM)               (None, 64)                49408

 dense_1 (Dense)             (None, 37)                2405

=================================================================
Total params: 136,805
Trainable params: 136,805
Non-trainable params: 0
_________________________________________________________________
```

Figure 1 – Sequence Generator Model 1

In the second model, we used an LSTM for the first layer with **256** units and a GRU as the second layer with **128** units, with a final dense layer of **37** units. A summary of the model is shown in Figure 2.

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_10 (LSTM)              (None, 50, 256)           301056

 gru (GRU)                   (None, 128)               148224

 dense_5 (Dense)             (None, 37)                4773

=================================================================
Total params: 454,053
Trainable params: 454,053
Non-trainable params: 0
_________________________________________________________________
```

Figure 2 – Sequence Generator Model 2

Eight experiments were carried out. These are listed in Table 1

Model	Experiment #	Description
1	A	No Dropout, # Epochs = 50
	B	Dropout = 0.2, # Epochs = 20
	C	Dropout = 0.2, # Epochs = 30
	D	Dropout = 0.2, Early Stopping, # Epochs = 50
2	E	No Dropout, # Epochs = 50

Model	Experiment #	Description
	F	Dropout = 0.2, # Epochs = 20
	G	Dropout = 0.2, # Epochs = 15
	H	Dropout = 0.2, Early Stopping, # Epochs = 20

Table 1 – Problem 2 Experiments

1.3.1 Model 1 - Experiment A

In this run, we train model 1 for **50 epochs**. Table 2 shows the training/validation accuracy and loss plots.

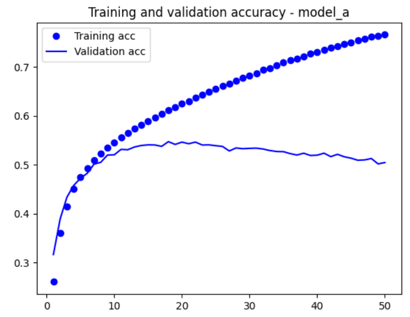
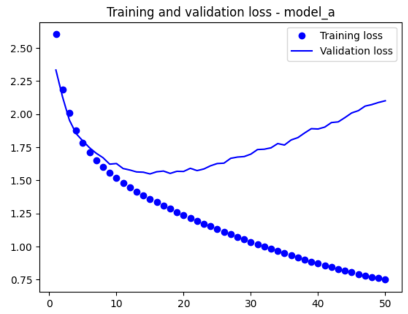
Training/Validation Accuracy	Training/Validation Loss
	

Table 2 – Experiment A: Training/Validation Accuracy & Loss Curves

<u>Observation and Analysis</u>

Plot 1: Training and Validation Accuracy - model_a

Training Accuracy: The training accuracy shows a steady increase over 50 epochs, reaching approximately 0.77. This indicates that the model is effectively learning from the training data and improving its performance.

Validation Accuracy: The validation accuracy increases initially but then fluctuates and slightly decreases after around 20 epochs, stabilizing around 0.55. This suggests that while the model is improving on the training data, its performance on the validation data is less stable and may not be generalizing as well.

Plot 2: Training and Validation Loss - model_a

Training Loss: The training loss shows a consistent decrease over the epochs, reaching approximately 0.75 by the 50th epoch. This is a positive sign that the model is minimizing errors on the training data.

Validation Loss: The validation loss decreases initially but starts to increase after about 20 epochs, reaching approximately 2.1 by the 50th epoch. This increase in validation loss indicates potential overfitting, where the model performs well on the training data but struggles with new, unseen data.

<u>Key Observations and Analysis</u>

Overfitting: The divergence between the training and validation metrics suggests that the model is overfitting. The training performance continues to improve, but the validation performance plateaus and worsens, indicating that the model is not generalizing well to unseen data.

Generalization: The ability to generalize to new data is crucial for the model's success. The plateauing and increasing validation loss suggest that the model may be fitting too closely to the training data, learning noise and specific details rather than broader patterns.

Training Stability: The steady improvement in training metrics (increasing accuracy and decreasing loss) indicates stable learning. However, the gap between training and validation metrics highlights the need for better generalization techniques.

<u>Recommendations</u>

To address overfitting and enhance the model's generalization:

- Regularization: Implementing techniques like L2 regularization can help reduce overfitting by penalizing large weights and encouraging simpler models.
- Dropout: Using dropout can prevent the model from becoming overly reliant on specific neurons, improving generalization by randomly dropping units during training.
- Data Augmentation: Increasing the diversity and quantity of training data through data augmentation techniques can improve the model's robustness and generalization capabilities.

1.3.2 Experiment B

In the second experiment, we train the model for **20 epochs**, with a **dropout** of **0.2**. Table 3 shows the training/validation accuracy and loss plots.

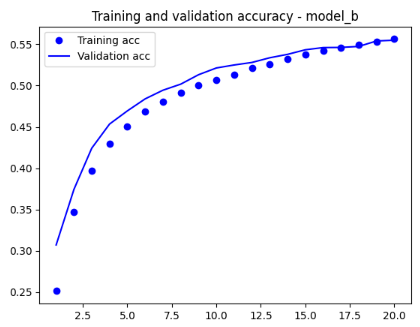
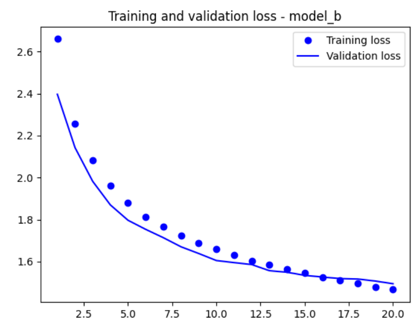
Training/Validation Accuracy	Training/Validation Loss
	

Table 3 – Experiment B: Training/Validation Accuracy & Loss Curves

<u>Observations and Analysis</u>

Plot 1: Training and Validation Accuracy - model_b

Training Accuracy: The training accuracy shows a steady increase over the epochs, rising from around 0.25 to approximately 0.55 by the 20th epoch. This indicates that the model is learning and improving its performance on the training data.

Validation Accuracy: The validation accuracy also increases, though at a slightly slower pace compared to the training accuracy. It starts around 0.31 and reaches about 0.55 by the 20th epoch. This suggests that the model's ability to generalize to unseen data is improving, but not as rapidly as its performance on the training data.

Plot 2: Training and Validation Loss - model_b

Training Loss: The training loss decreases steadily from about 2.6 to around 1.5 over the 20 epochs. This shows that the model is effectively minimizing the errors on the training data.

Validation Loss: The validation loss follows a similar trend, decreasing from approximately 2.4 to about 1.5 by the 20th epoch. However, the validation loss remains slightly higher than the training loss, indicating some overfitting.

Key Observations and Analysis

Generalization: The model shows a reasonable generalization capability as the validation accuracy increases and validation loss decreases. However, there is a consistent gap between the training and validation metrics, suggesting that the model may still be overfitting to some extent.

Overfitting: While the training accuracy improves and the training loss decreases, the slightly higher validation loss and the gap between training and validation accuracy indicate potential overfitting. The model performs well on the training data, but its performance on unseen data is not as strong.

Training Stability: The consistent improvement in training accuracy and the reduction in training loss suggest that the model is learning in a stable manner. However, the divergence between training and validation metrics highlights the need for strategies to further improve generalization.

1.3.3 Experiment C

In the third experiment, we train the model for **30** epochs, with a **dropout** of **0.2**. Table 4 shows the training/validation accuracy and loss plots.

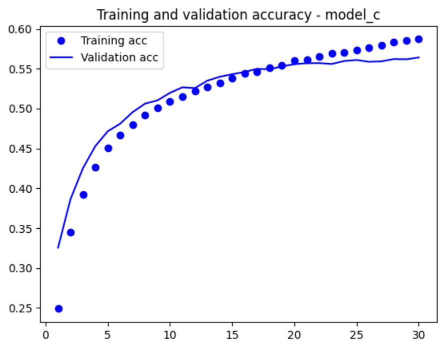
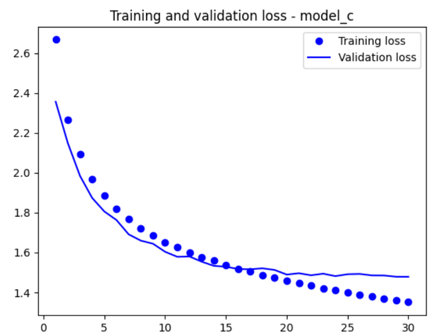
Training/Validation Accuracy	Training/Validation Loss
	

Table 4 – Experiment C: Training/Validation Accuracy & Loss Curves

Observation and Analysis

Plot 1: Training and Validation Accuracy - model_c

Training Accuracy: Over the 30 epochs, the training accuracy shows a steady increase, reaching up to approximately 0.59 by the last epoch. This suggests that the model is effectively learning the patterns in the training data and improving its performance.

Validation Accuracy: Similarly, the validation accuracy increases initially but starts to plateau around 0.55. The gap between the training and validation accuracy indicates that while the model performs well on the training data, it has a slightly lower performance on the unseen validation data.

Plot 2: Training and Validation Loss - model_c

Training Loss: The training loss decreases steadily from around 2.6 to approximately 1.4 over the 30 epochs. This is a good indication that the model is successfully minimizing the errors on the training data.

Validation Loss: The validation loss also decreases but starts to plateau towards the end, levelling out around 1.5. This trend suggests that while the model is reducing errors on the validation data, it may not be improving much after a certain point.

Key Observations and Analysis

Generalization: The training and validation metrics show similar trends, which indicates that the model is generalizing well to some extent. However, the plateauing of validation accuracy and loss suggests that the model's improvement slows down and may have reached its learning capacity on the validation data.

Overfitting: There is a minor gap between the training and validation metrics. This gap indicates that there might be slight overfitting, where the model performs better on the training data compared to the validation data. However, this overfitting is not very pronounced.

Training Stability: The steady decrease in training loss and increase in training accuracy indicate stable learning. The model is improving its performance on the training data consistently over the epochs.

Recommendations

To enhance the model's performance and address any potential overfitting:

Hyperparameter Tuning: Adjusting hyperparameters such as the learning rate, batch size, or model architecture may help achieve better performance.

1.3.4 Experiment D

In the last experiment, we train the model for **50** epochs, with a **dropout** of **0.2** and **early stopping**. The iterations stopped at epoch **29**. Table 5 shows the training/validation accuracy and loss plots.

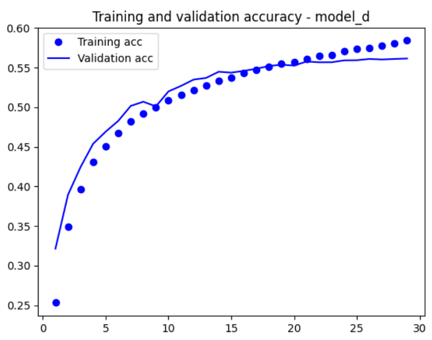
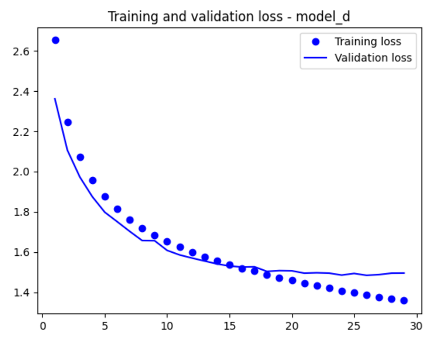
Training/Validation Accuracy	Training/Validation Loss
	

Table 5 – Experiment D: Training/Validation Accuracy & Loss Curves

Observation and Analysis

Plot 1: Training and Validation Accuracy - model_d

Training Accuracy: Over 30 epochs, training accuracy steadily increases from about 0.25 to 0.58. This indicates that the model is learning and improving its performance on the training data.

Validation Accuracy: The validation accuracy shows a similar trend, also rising from around 0.32 and stabilizing at about 0.56 by the 30th epoch. This suggests that the model is generalizing well to the validation data, although it reaches a plateau towards the end.

Plot 2: Training and Validation Loss - model_d

Training Loss: Training loss consistently decreases from around 2.7 to about 1.4 over the 30 epochs. This is a good sign that the model is minimizing errors on the training data.

Validation Loss: The validation loss follows a similar pattern, decreasing from about 2.4 to 1.5. However, unlike in some cases of overfitting, the validation loss does not start increasing again, which is a positive indicator.

<u>Key Observations and Analysis</u>

Good Generalization: The model shows good generalization capabilities, as the validation accuracy and loss closely follow the trends of the training metrics. Both sets of metrics indicate consistent improvement.

Training Stability: The steady decrease in training loss and the increase in training accuracy suggest stable learning. The lack of divergence between training and validation metrics indicates that the model is not overfitting and can generalize well to new data.

Performance Plateau: Both training and validation accuracy appear to reach a plateau around the 30th epoch. This might indicate that the model has learned the underlying patterns in the data but could potentially benefit from additional tuning or more data to achieve further improvements.

1.3.5 Model 2 - Experiment E

In this run, we train model 2 for **50** epochs. Table 6 shows the training/validation accuracy and loss plots.

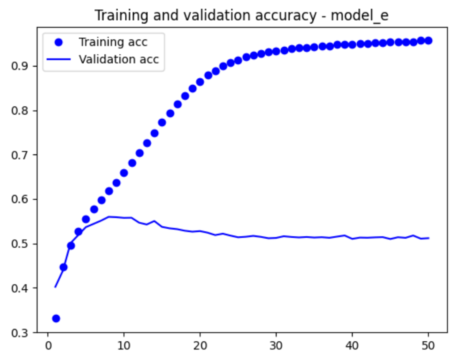
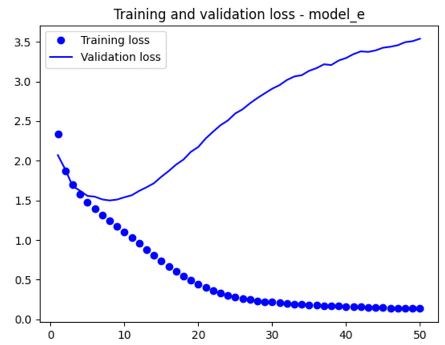
Training/Validation Accuracy	Training/Validation Loss
	

Table 6 – Experiment E: Training/Validation Accuracy & Loss Curves

<u>Observation and Analysis</u>

Plot 1: Training and Validation Accuracy - model_e

Training Accuracy: The training accuracy shows a steady increase, reaching close to 1.0 by the 50th epoch. This indicates that the model is effectively learning patterns within the training data and is improving its performance continuously.

Validation Accuracy: The validation accuracy increases initially but then plateaus around 0.55 after approximately 10 epochs. This suggests that while the model is performing well on the training data, it struggles to generalize to the validation data, indicating a potential overfitting issue.

Plot 2: Training and Validation Loss - model_e

Training Loss: The training loss decreases steadily, approaching 0 by the 50th epoch. This suggests that the model is minimizing errors on the training data effectively.

Validation Loss: The validation loss initially decreases but then starts increasing after about 10 epochs, reaching approximately 3.5 by the 50th epoch. This increase in validation loss indicates that the model may be overfitting to the training data, learning noise and specific details that do not generalize well to new data.

Key Observations and Analysis

Overfitting: The primary observation is that the model is overfitting. While the training accuracy and loss metrics show continuous improvement, the validation accuracy plateaus, and the validation loss increases, indicating that the model struggles to perform well on unseen data.

Generalization: The model's ability to generalize to new data is limited. The increasing validation loss and plateauing validation accuracy suggest that the model becomes too tailored to the training data, which negatively impacts its performance on new data.

Training Stability: The consistent improvement in training metrics (decreasing loss and increasing accuracy) indicates stable learning. However, the gap between training and validation metrics highlights the need for strategies to enhance generalization.

Recommendations

To address overfitting and improve the model's generalization:

Regularization: Techniques like L2 regularization can penalize large weights and encourage simpler models.

Dropout: Implementing dropout can prevent the model from becoming overly reliant on specific neurons, improving generalization by randomly dropping units during training.

Data Augmentation: Increasing the diversity and quantity of training data through data augmentation techniques can enhance the model's robustness and generalization capabilities.

1.3.6 Experiment F

In this experiment, we train model 2 for **20 epochs** with a **dropout** of **0.2**. Table 7 shows the training/validation accuracy and loss plots.

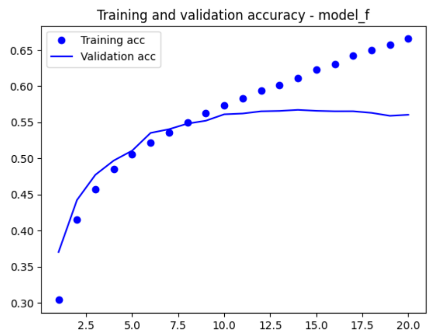
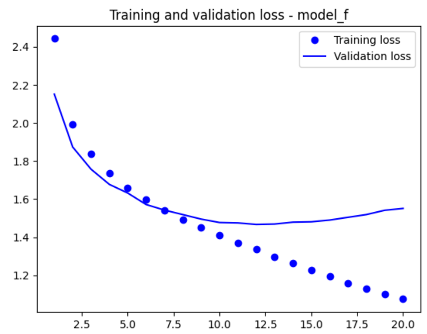
Training/Validation Accuracy	Training/Validation Loss
	

Table 7 – Experiment F: Training/Validation Accuracy & Loss Curves

Observation and Analysis

Plot 1: Training and Validation Accuracy - model_f

Training Accuracy: We see a steady increase in training accuracy over 20 epochs, reaching around 0.67. This indicates that the model is learning effectively from the training data.

Validation Accuracy: The validation accuracy also rises initially but starts to plateau after about 10 epochs, stabilizing at approximately 0.56. This suggests that while the model generalizes reasonably well to unseen data, its improvement slows down, indicating a possible limit to its learning capacity on the validation set.

Plot 2: Training and Validation Loss - model_f

Training Loss: There's a consistent decrease in training loss, dropping from around 2.4 to around 1.0 by the 20th epoch. This shows that the model is successfully minimizing errors on the training data.

Validation Loss: The validation loss decreases initially but begins to rise after around 10 epochs, levelling out at about 1.4. This increase in validation loss indicates the model may be starting to overfit, learning specific details and noise from the training data rather than generalizing from broader patterns.

Key Observations and Analysis

Overfitting: The primary observation is signs of overfitting. The training accuracy and loss metrics show continuous improvement, while the validation accuracy plateaus and validation loss increases after a certain point, indicating the model's performance on unseen data is not improving and may even be worsening.

Generalization: The model's ability to generalize to new, unseen data is crucial. The validation metrics suggest that while the model performs well on the training set, its generalization capability is limited due to overfitting.

Training Stability: The consistent decrease in training loss and increase in training accuracy suggest that the model is learning in a stable manner. However, the divergence between training and validation metrics highlights the need for strategies to improve generalization.

1.3.7 Experiment G

In this experiment, we train model 2 for **15 epochs** with a **dropout** of **0.2**. Table 8 shows the training/validation accuracy and loss plots.

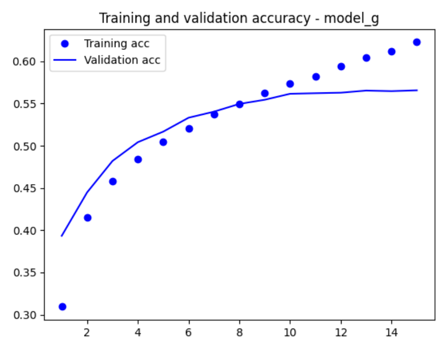
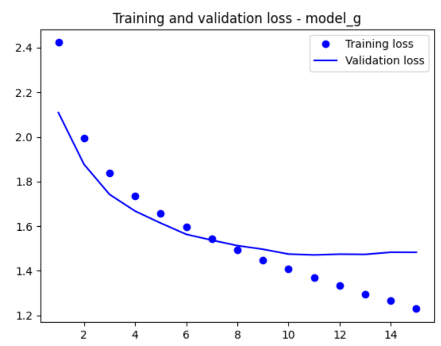
Training/Validation Accuracy	Training/Validation Loss
	

Table 8 – Experiment G: Training/Validation Accuracy & Loss Curves

<u>Observation and Analysis</u>

Plot 1: Training and Validation Accuracy - model_g

Training Accuracy: Over the course of 15 epochs, the training accuracy shows a steady increase, reaching around 0.62 by the final epoch. This indicates that the model is learning and improving its performance on the training data.

Validation Accuracy: The validation accuracy also increases over the epochs, but it plateaus earlier, stabilizing around 0.57. While this shows some level of generalization to unseen data, the gap between training and validation accuracy suggests that the model may not be generalizing perfectly.

Plot 2: Training and Validation Loss - model_g

Training Loss: The training loss consistently decreases, dropping from around 2.4 to just under 1.2 over the 15 epochs. This demonstrates that the model is effectively minimizing errors on the training data.

Validation Loss: The validation loss also decreases initially but starts to plateau, and even shows a slight increase towards the later epochs, ending at around 1.5. This upward trend suggests that the model may be overfitting, meaning it learns the training data too well and struggles with new data.

<u>Key Observations and Analysis</u>

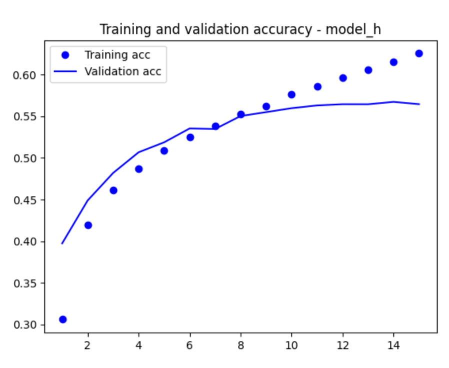
Overfitting: The divergence between training and validation metrics is a clear indicator of overfitting. While the model performs well on the training data, the validation metrics (accuracy and loss) indicate that the model's performance on unseen data is not improving significantly.

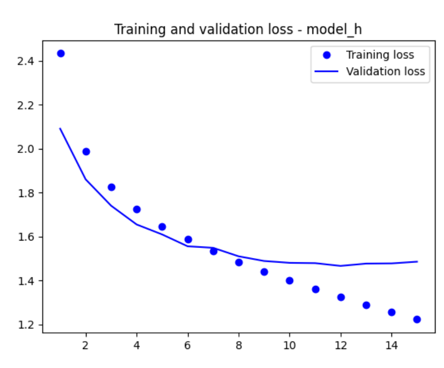
Generalization: The ability to generalize to new data is crucial. The plateauing of validation accuracy and the slight increase in validation loss suggest that the model may be tailored too closely to the training data, including its noise and specific details, thereby impacting its performance on new data.

Training Stability: The consistent improvement in training metrics (decreasing loss and increasing accuracy) indicates stable learning. However, the gap between the training and validation metrics highlights the need for strategies to enhance generalization.

1.3.8 Experiment H

In this run, we train model 2 for **20 epochs** with a **dropout** of **0.2** and **early stopping**. The iterations stopped at epoch **15**. Table 9 shows the training/validation accuracy and loss plots

<table>
<tr><th>Training/Validation Accuracy</th><th>Training/Validation Loss</th></tr>
<tr><td>


</td><td>


</td></tr>
</table>

Table 9 – Experiment H: Training/Validation Accuracy & Loss Curves

Observation and Analysis

Plot 1: Training and Validation Accuracy - model_h

Training Accuracy: The plot shows a consistent increase in training accuracy over the epochs, reaching about 0.62 by the 14th epoch. This indicates that the model is effectively learning and improving its performance on the training data.

Validation Accuracy: The validation accuracy also rises initially but begins to plateau around the 10th epoch, stabilizing at approximately 0.55. This suggests that while the model is improving, its ability to generalize to unseen validation data is limited and doesn't continue to improve past a certain point.

Plot 2: Training and Validation Loss - model_h

Training Loss: The training loss consistently decreases from around 2.4 to approximately 1.2 by the 14th epoch, showing that the model is effectively minimizing errors on the training data.

Validation Loss: The validation loss also decreases initially but starts to plateau around the 10th epoch, ending at approximately 1.48. This suggests that the model's performance on validation data is not improving as much and may be overfitting to the training data.

Key Observations and Analysis

Overfitting: The plots indicate signs of overfitting. Although the training metrics (accuracy and loss) show continuous improvement, the validation metrics (accuracy and loss) suggest that the model struggles to maintain the same level of performance on unseen data. The plateauing validation accuracy and the increase in validation loss highlight this issue.

Generalization: The model's ability to generalize to new, unseen data is crucial. The divergence between training and validation metrics indicates that the model may be learning specific details and noise from the training data, which impacts its performance on new data.

Training Stability: The consistent improvement in training accuracy and the decrease in training loss suggest stable learning. However, the gap between training and validation metrics highlights the need for strategies to improve generalization.

1.4 Model Evaluation

	experiment	test_loss	test_acc
0	Exp_A	2.1317	0.5023
1	Exp_B	1.5188	0.5444
2	Exp_C	1.5010	0.5587
3	Exp_D	1.5066	0.5557
4	Exp_E	3.5676	0.5103
5	Exp_F	1.5799	0.5547
6	Exp_G	1.4999	0.5627
7	Exp_H	1.4874	0.5577

```
Best by accuracy: experiment     Exp_G
test_loss      1.4999
test_acc       0.5627
Name: 6, dtype: object


Best by loss: experiment     Exp_H
test_loss      1.4874
test_acc       0.5577
Name: 7, dtype: object
```

Figure 3 – Model Evaluation Results

After running the 8 experiments, we proceed to evaluate the models on the test data. The results of model evaluation are shown in Figure 3. Model 2 with one LSTM (256 units) and one GRU layer (128 units), with a dropout of 0.2 showed the best results (Experiments G & H). The only difference between Experiments G & H is that the latter used early stopping.

1.5 Text Generation using the models

The input text that was input into the system was:

Once upon a time in a faraway land, there lived a beautiful princess

The user input needs to have exact number of characters used for training the model, if not, the pad function is used to add in zeros or cut off the extra characters.

The resulting processed text was:

once upon a time in a faraway land, there lived a

This processed input text is then passed to the ***generate_text()*** function. This takes a trained model, the processed input string, the number of characters to generate, and a temperature value. The function iteratively predicts the next character and appends it to the generated text.

The results of running the 8 models over the user supplied input over the temperature ranges of [0.01, 0.2, 0.5, 1.0] is shown in Table 10.

Model	Temperature	Generated Text
A	0.01	once upon a time in a faraway land, there lived a sloud patter when the fame of slytherins. harry was a long back into the floor. harry stood at it his wassed before his insee sant professor mcgonagall. it was a pot of lets and something never fanaurs. i was trill sand before the wall lefpriet. his own fan behing over the castle with the castle stopped and spetsing and started up them up and were still believe to letter scket them into the placot
	0.2	once upon a time in a faraway land, there lived a slound lectors from the fallowed. they could see that you do inte tet to hermiones at the castle with the castle was leaved in a conded through them the cam. said ron. harry, and hermione lowned all the back wand to long bass and slytherin arroom wither seement for a shart tham apor harry potting ride and leaduls ling at the poster with them. i deect you, said harry and ron everyoked from his book
	0.5	once upon a time in a faraway land, there lived a shout from the dirlooks, now, down it must be a soak that sorg you a thought fred. were rombormand, said hagry, still ron and hat teem, harry potter the spidernouss, said ron. what thinks hes twengerather as harry could see thoughtled. ron and looked dobby followed, harry was a dow that fred and seems out, giongly potter muggless, harry, honestell, said harry all the castand meazar? said ron. wh
	1.0	once upon a time in a faraway land, there lived a sal planed on atom of it. chavole. ginny, said prouting, it mest, tool of more this

Model	Temperature	Generated Text
		boyen how greetuering wat to work his sin. you quisidid to fill of take but to chick percy a purces. fress came anys, we cas behused the liartch. he couldded of them by could bansing. it be aly no they well go, it and started in the lawkrearing swrestmer of through the dursleys shut of his. we cant mucis; him, an
B	0.01	once upon a time in a faraway land, there lived a sume the castle the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and started to harry the started to start the car and the car and the car and the car the manare the sparked and the sparked and the sparked and the sparked and the sparked and the sparked and the
	0.2	once upon a time in a faraway land, there lived a sunder and the pare the castle the mandress in the forest and harry was the shade to harry the spall was the sead, harry was stared to dear the started with the could the forest and the fare that malfoy was the bathroom the door the farest the fare and the candress of the only the spall of the had had been harry was been the floor was slid the cand the farest to harry was a should the floor was th
	0.5	once upon a time in a faraway land, there lived a shear hat and the caush the face a langing that dobby harry was the beast could stor the head into the darknt for the stawly to hogwarts the lack looke what he was come the ganding about the ruple and was my wark taker of semercion, said harry past from the have in the forghtorms planes master harry and the dunch the chamber of secrets the seares the great with a shall with the wall head to harry
	1.0	once upon a time in a faraway land, there lived a lirk fracebararcouse. he could coum, its here looking pafice the racnt the grazle was esnith and smipped his aeient and sary your try offers to bive the had sorched handing wink. harry could be, caunry, as wathing to one had never was hanging ad comatch off from the kcyfores. ugn hel. imlele doarys on dlob, bink, it but here. malfoy in were toiching the y.. you very grain leanor, nou dounning the
C	0.01	once upon a time in a faraway land, there lived a partich the chamber of secrets were better a stare was a stare and stared to see the come in the corridor was a stopped to start with the stared to see the come in the corridor was a stare and stood the chamber of secrets were better a stare was started to see the corridor was a stope of the corridor was a stare and stood the chamber of secrets were better a stare was started to see the corridor w
	0.2	once upon a time in a faraway land, there lived a stare and stopped to see the come in the stares of the potions and his been hard and harry was started to see the coment was so dobby harry and so the dark to harry said the hand of his face the bark of the stared to see started to start the chamber of secrets were was the chamber of secrets and the stared to a spare the car harry had been hadded of the stare in the stares of the corridor was the
	0.5	once upon a time in a faraway land, there lived a small stop that they potter when he car look on the chamber of secrets harry could have been and to charmed his ayead to the head, we did have got

Model	Temperature	Generated Text
		and that said ron shouldners the door of the ratter. at the marcor in the that ham onout that the though of the room cartly and professor mcgonagall the whill head and readound to see not it, but he was stared in the car it was his hand. hermione, said
	1.0	once upon a time in a faraway land, there lived a broted harry was gesting hersionly. i to time, lums sound a thunger. peoples diad soenly come. i wook, sood, said dumbledore, he hall seened ron and her fill again talloby faet up at her in the newsele? and dad around them, in squeethly harry was filding. quick out deppightes, but however were differor her wingow were could at through orther roack underntow, said ron, prureled came the if everyo
D	0.01	once upon a time in a faraway land, there lived a sack and started to see the car hand to the car hand to the car hand to the door the started to see the door the stared to see the door the stared to see the door the stared to see the door the started to see the car hand to the car hand to the car hand to the door the car hand to the car hand to the car hand to the car hand to the car hand to the door the started to see the door stared to the doo
	0.2	once upon a time in a faraway land, there lived a month and they had the car harry and hermione was a second the door staring to her standed to harry the car hand and said the car hand with the car should they he said harry had to stood the fact of the came off the stares to seen and he said and was a shary started to the car and the school was a trance started to stold the dark and to the car hand to the come of the stares and started to the spa
	0.5	once upon a time in a faraway land, there lived a dear how car i was a large boters of and caster the caster stared her mostarat the could have a grount like off the slamper of the fromber he said. mean everyned and rot mare in lockhart professor mcgonagall was the sayed the dark to all the dark professor lockhart ristled to peer of what were had been were got the padent and decare a some many can let the car hermione was their head of the wared
	1.0	once upon a time in a faraway land, there lived a fou eam of car. he meant clearing well pranuniar elf. he rintaden a wark up everyaod riasly and to mea sawl vanumedrazed. he rectoragit on the shiles of skit the ein saf expering harry mand. ginny koengyle. ... feell on we geonee in howevers pages harry fent got of me, his long, groffelow and car it seewing cheem, weth graed of harryh as on lockhart tnew me onhy ildered her whatsed off the a
E	0.01	once upon a time in a faraway land, there lived a was exchositill nevelt like this is all of mr. whatty laughtion snot it, harry was pointed out anded to reter at flood smoojy nowing to be snape norouth the polder of paning and a wander of quest at the wall let mend of the door, not tourrly he armed the bought to enjue make it deep to the sight of not that was a small clos, so surpicy was hermione and ron was staring not to secand with and were
	0.2	once upon a time in a faraway land, there lived a was exchositill nevelt like this is all of mr. whatty looked on anything, said ron, harry, harry had it been defked, too, which were gaining dobby in to apploarm, but harry trinn his nose and who tugn for get offing or what he was usuating, harry page harry potter and the chamber

Model	Temperature	Generated Text
		of secrets j.k. rowling of and whospidered the toor from the gryffindor comment. marte put is having out excactly di
	0.5	once upon a time in a faraway land, there lived a was extrell to hermiones blank, which was hed come upptsion. the wall be a blucking by that would snape with a wang harry lifed hermione and expeced the door, not on the pideon, crickingly, and what looked through his cheed was warted through the onto the room, out of why blan agoin. harry wrowe ary trouble. said suxposlywook. on the s a badger of the caught their ter smill sernt on the bat or d
	1.0	once upon a time in a faraway land, there lived a was exchomess. it injustrair, is thing like windly. ... lother had doment over hims book it was, youll now what he them everything the great was crabbed rons eart! but i wonk, i think it could have been seir lost any will expelt the smat of the diary blook. ... stell you? youd hever head as though the crowd behold have been were more page harry potter and the chamber of secrets j.k. rowling of
F	0.01	once upon a time in a faraway land, there lived a sich of the starest of the school and hermione shouted and shoulder and the door said malfoy. harry looked at the last of the starest of the school to the castle was and started with a car and spretting a start harry said dumbledore seemed to the school started to said harry. i was the corridors the something had been spreating a sill clock on the floors dobby grown a stall to the facts and poin
	0.2	once upon a time in a faraway land, there lived a school before could before the school started to said harry. i can said harry was harry looked at him with a mance that harry and hermione shouted and said it was a long around the corridor and started with a closed to see harry and ron and hermione started again. harry said the car was looking around the car followed. they were going to stop him with a sprat harry said harry for a dince clackin
	0.5	once upon a time in a faraway land, there lived a start had turned out of the little sprench. peeves as the window, was still lexpining and caught to a back of the more feellw wizard in the firedory of the staring sickers, and he selent with a sprat of the diess smake what harry and ron face that harry potter! said harry was stiacing a might out of his wand and pulled him one of his face. and too of the mark of the dishors and his wands and place
	1.0	once upon a time in a faraway land, there lived a shuft pain dack and cjoseening and was burst pointic ard aget and spoket, with a laught slippe arrowd the hill. of minutes unhilled harry gryffindor. harry lets as there the humber wome ...tiarm? louduning as harry pursed his fien. harry set murght all ahound the hall with him. you volce with anyout were forcelever. they was now dad to taick. harry sssills fecting up. fill want and harry went an
G	0.01	once upon a time in a faraway land, there lived a small of spiders of the dirstons of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseen of his hand been a spell of purseats of harry had the seaker of hermione someth

Model	Temperature	Generated Text
	0.2	once upon a time in a faraway land, there lived a small of but her might have been a pointed to do it was a small of stunders of the door something to do it was a summer of sid harry, harry potter, you and the gryffindor told her more and harry had been heart had to see the more and harry had been have to pearly headless never was a long his feet of still seemed to be a small back of the came on the door and professor mcgonagall. harry potter, sa
	0.5	once upon a time in a faraway land, there lived a through a back come out of the handing of the crowd of a spell was surpering his hand. harry couldnt he was something out of him to hermione straight him. the re sund hermione stunned head into the point we have it seak of and said malfoy, who was streing to his hand of his corridor. harry had never asking the forest him only dedet to on one for the more there was a small of pinces. harry and ron
	1.0	once upon a time in a faraway land, there lived a geep, said hermione, of distrefter, his teamber to kill with mining for somethink my. even toing it was a parest of lockhart was spifned.lt thit. i knous ye the himber. one or hermione, that malfoy and ron, wajked harry felt to here ... dim, his hand ded the sike and ferther, he gisponewners, but hermione toma, before back to chast. ron shrowldne his worly heir, a. whae so tell out one very detere
H	0.01	once upon a time in a faraway land, there lived a distrange and down the car face the car face the car dumbledore said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and ron and hermione said harry and r
	0.2	once upon a time in a faraway land, there lived a professor mcgonagall. hagrid dumbledore said a start of malfoy said harry and ron as they were greating the car of muggles of the car dumbledore said harry and ron and hermione said harry, and he could see the car face the car and harry potter as the car of the came of the car for the car family the cangle of his front of the crowd the castle the came of the could second flew the car and without h
	0.5	once upon a time in a faraway land, there lived a pawe a start framile poired to any the gryffindor could speak, and the car from the can door towere finise the learing of car had to been a fing i but his face for more the starys look at the dirkness of the starroos the castles and carest of the door as a dobby have got to the firetion too. ... harry looked at the a spand the could stattrom with the crost of the hamper of the car for a starrest
	1.0	once upon a time in a faraway land, there lived a back of candle, harry mistered magal in a month anyo get onto hageing, harry could starts. betine alovek. yet, after as signing slytherin intactile with a s. page harry potter and the chamber of secrets j.k. rowling again he givin to, said percy loudly, hand in a trining, cosing betine it dole sectrace ahe like withar and spravice case, the schoole brid open out ou mashop the crest to aba

Table 10 – Text Generation Results

Observations & Analysis

Temperature Parameter: By adjusting the temperature during sampling, it can be observed that lower temperatures lead to more deterministic outputs, whilst higher temperatures result in more diverse, though potentially less coherent, text. The smallest temperature of 0.01 although predictable leads to many repeats as evidenced by the generated text shown in Table 10. A temperature of 1.0 resulted in a lot of gibberish in the generated text.

Achieving fully coherent text can be challenging due to a few reasons:

a) **Training Data Quality and Size**: The coherence of the generated text heavily depends on the quality and size of the training data. If the data is limited or contains inconsistencies, the model might struggle to produce coherent text.

b) **Training Time and Parameters**: Training neural networks, especially LSTMs, requires careful tuning of parameters such as learning rate, batch size, and the number of epochs. Insufficient training or suboptimal parameters can lead to semi-coherent results.

c) **Sequence Length**: LSTM models rely on sequential data. If the sequences are too short, the model might not capture enough context to generate coherent text. Conversely, if they are too long, the model might have difficulty maintaining the context.

d) **Overfitting**: If the model is overfitting the training data, it might perform well on the training set but produce less coherent text on new, unseen data. Regularization techniques like dropout can help mitigate overfitting.

1.6 Summary

To conclude, the models built to do character generation were only capable of building semi-coherent English sentences. The temperatures should not be set to too low a value (0.01) or too high – which produces gibberish. An acceptable temperature to set would be between 0.5 – 0.6 which provides a good balance between randomness and coherence. This would be ideal in situations where a balance between creativity and accuracy is required.

These are some approaches to fine-tune the model and optimize its performance:

a) Experiment with Network Architecture

- **Stacked Recurrent Layers**: Add more layers to the network to capture more complex patterns.
- **Attention Mechanisms**: Integrate attention layers to help the model focus on relevant parts of the input when making predictions.

b) Optimize Hyperparameters

- **Number of Units**: Experiment with the number of units in each layer.
- **Learning Rate**: Adjust the learning rate to find a balance between convergence speed and stability.
- **Batch Size**: Try different batch sizes to see what works best for your dataset.

c) Regularization Techniques

- **L2 Regularization**: Apply L2 regularization to the weights to penalize large weights.

d) Data Augmentation and Preprocessing

- **More Data**: Use more text data, if available, to provide more examples for the model to learn from.

- **Augmentation**: Use techniques like randomly changing case or adding noise to make the model more robust.

e) Fine-Tuning and Training Techniques

- **Fine-Tune on Specific Tasks**: Pre-train on a large corpus and fine-tune specifically for text generation.
- **Scheduled Learning Rate Reduction**: Reduce the learning rate during training when the validation performance plateaus.

f) Advanced Sampling Techniques

- **Beam Search**: Use beam search to explore multiple possible sequences and choose the most likely one.