**Name**: Chia Bing Xuan

**Matriculation Number**: A0259419R

**Title**: DSA4213 Assignment 2 – Small Language Models Exploration

1. **Introduction**

In this assignment, I conducted experiments on two types of models – vanilla Recurrent Neural Networks (RNNs) and Long Short-Term Memory RNNs (LSTMs). With the use of a selected corpus, RNNs and LSTMs were built and trained for language modelling.

1. **Explanation of Models**
   1. **RNNs**

Note that RNNs can process input sequences of any length. In the context of language modelling, suppose we have an input sequence of textual tokens, where is the sequence length. At each sequential time step , the corresponding token is fed into the RNN. It is first converted into a word embedding with the transformation matrix

The RNN possesses a hidden state that will be updated during each time step. During time step , the hidden state from the previous step, , will be updated using the word embedding to yield the new hidden state

where is the sigmoid function and , ,are learnable parameters. Finally, the hidden state of the current time step is used to output a probability distribution over all distinct tokens in the vocabulary :

where andare learnable parameters in the RNN. To predict the next token, it can be chosen by selecting the token corresponding to the highest probability. Alternatively, the next token can be sampled from the output probability distribution.

Note that the same weights , will be repeatedly applied on every time step. They will only be updated through gradient descent after all the input tokens in the sequence have been processed. Hence, there is symmetry in how these input tokens are being processed.

* 1. **LSTMs**

LSTMs are RNNs that aim to alleviate the vanishing gradient problem – an issue which makes it difficult for vanilla RNNs to learn long-range dependencies and preserve information over many timesteps.

Suppose we have completed the previous time step . The LSTM has a hidden state and a cell state , the latter of which can store long-term information. To update the hidden state and cell state for the current time step , we first determine three gates – forget gate , input gate and output gate . We also determine the new cell content that can be written to the cell in this time step, . These vectors are dynamically obtained based on the current word embedding :

where the ’s, ’s and ’s are learnable parameters. To determine the new cell state , the forget gate selectively removes some content from the previous cell state , while the input gate selectively writes some new cell content from :

Finally, the output gate selectively reads some content from the updated cell state , so as to obtain the updated hidden state for the current time step:

1. **Methodology**
   1. **Selection of Corpus and Data Processing**

I selected a corpus of Reuters news documents offered by the NLTK library, which consists of 10788 news documents. Owing to limitations in computational power, only a subset of the dataset was used (2500 documents). Firstly, a train-validation-test split of 80/10/10 was applied to this subset. This enables us to obtain a training set of 2000 documents, a validation set of 250 documents and a testing set of 250 documents. For each of these documents, we then applied a round of data pre-processing:

1. Tokenise the document. I experimented with two methods of tokenisation – *word tokenisation* using torchtext’s basic English tokeniser, and *subword tokenisation* using a base BigBird transformer model (based on SentencePiece).
2. For each token produced, convert it to lowercase
3. Add <unk> (unknown), <bos> (beginning of sentence) and <eos> (end of sentence) tokens
4. Numericalise the tokens by converting them to unique integer IDs
   1. **Model Training**

Subsequently, I trained RNNs and LSTMs for language modelling, on both the word tokens and subword tokens from the training set. For this investigation, tokenisation method and dropout were selected as the dependent variables. Across all the trials, the following were kept constant:

* Embedding size = 128 for the embedding layer of the RNN / LSTM
* Hidden size = 256 for the hidden state of the RNN / LSTM
* Number of layers = 2 for the RNN / LSTM unit
* Sequence length = 32 tokens. A small value was chosen, owing to computational constraints
* Number of epochs = 10
* Batch size = 32
* Optimiser = Adam
* Learning rate =
* Maximum norm for gradient clipping = 1.0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Trial** | **Model** | **Tokenisation Method** | **Dropout (between RNN / LSTM layers)** | **Training Time** | **Final Validation Cross Entropy Loss** | **Final Validation Perplexity** |
| 1 | RNN | Word | 0.0 | 23 min 24 s | 9.2678 | 10591.1113 |
| 2 | RNN | Word | 0.2 | 24 min 16 s | 7.4438 | 1709.1497 |
| 3 | RNN | Subword | 0.0 | 30 min 49 s | 7.8859 | 2659.3932 |
| 4 | RNN | Subword | 0.2 | 32 min 12 s | 6.2666 | 526.7011 |
| 5 | LSTM | Word | 0.0 | 23 min 5 s | 10.5475 | 38083.9244 |
| 6 | LSTM | Word | 0.2 | 23 min 35 s | 8.5566 | 5201.0926 |
| 7 | LSTM | Subword | 0.0 | 29 min 43 s | 9.1753 | 9655.6448 |
| 8 | LSTM | Subword | 0.2 | 30 min 24 s | 7.2038 | 1344.5489 |

A group of graphs with numbers

AI-generated content may be incorrect.

*For loss curves: Row 1 (left to right) – Trials 1, 2, 3 and 4. Row 2 (left to right) – Trials 5, 6, 7 and 8*

Across all 8 trials, the training loss tends to 0, while the validation loss increases with each epoch. Hence, each of the models has overfitted to the training data during the training process. This can be attributed to the fact that only a small subset of the Reuters dataset was used, which might be insufficient for model training. Since the training set is relatively small, the tokens in the validation set are more likely to be absent in the training vocabulary. Therefore, it is possible that the training set and validation set are significantly different from one another, which explains why the validation loss increases as the models fit to the training data. Trial 4 corresponds to the smallest validation loss and perplexity, displaying the smallest degree of overfitting.

1. **Evaluation** 
   1. **Quantitative Analysis (Cross Entropy Loss and Perplexity)**

Each of the trained models was evaluated on a testing dataset. Model performance was quantified using two metrics – cross entropy loss and perplexity.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial** | **Model** | **Tokenisation Method** | **Dropout** | **Test Cross Entropy Loss** | **Test Perplexity** |
| 1 | RNN | Word | 0.0 | 10.0254 | 22593.0692 |
| 2 | RNN | Word | 0.2 | 8.0771 | 3219.9619 |
| 3 | RNN | Subword | 0.0 | 8.5430 | 5130.6781 |
| 4 | RNN | Subword | 0.2 | **6.8204** | **916.3774** |
| 5 | LSTM | Word | 0.0 | 11.6870 | 119014.8032 |
| 6 | LSTM | Word | 0.2 | 9.4750 | 13029.5869 |
| 7 | LSTM | Subword | 0.0 | 10.0901 | 24103.5258 |
| 8 | LSTM | Subword | 0.2 | 7.9883 | 2946.2009 |

For the same model type and dropout value, subword tokenisation consistently achieves a better performance than word tokenisation. This is likely because subword tokenisation more robustly allows the models to reason about structures below the word level – such as common prefixes, suffixes and roots. This is useful when dealing with many variants of the same root word (eg. “help”, “helping”, and “helper”). With word tokenisation, the model might not be able to learn that these words are semantically similar, since they are treated as distinct tokens. With subword tokenisation, not only is the model able to learn a strong representation for the root word itself (“help”), but it can also learn the use of prefixes and suffixes (“ing” and “er”) by generalising to the structures of other words (eg. “singing” and “singer”). Furthermore, for rare words that would have otherwise been outside of the model’s vocabulary, subword tokenisation helps break them down into more common subwords that the model has seen before. Hence, the model is still able to infer the meanings of such words.

For the same model type and tokenisation method, the addition of dropout consistently improves performance. This is unsurprising as the random dropout of neurons prevents them from depending too heavily on one another, causing neuron co-adaptation to occur to a smaller extent. Hence, this reduces the likelihood of the model overfitting to the training data, allowing it to generalise to the testing data more effectively.

For the same tokenisation method and dropout value, vanilla RNNs appear to consistently outperform LSTMs. There are two possible reasons for this. Firstly, the size of the training dataset is rather small. Since LSTMs have more trainable parameters than RNNs, the higher model complexity means that LSTMs are more likely to overfit during training, leading to poorer generalisation to the testing data. Secondly, note that LSTMs are advantageous when it comes to preserving information over a long sequence of time steps. Since the context length selected (32) is rather short anyway, there is less of a need to capture, preserve and learn long-range dependencies. This suggests that the benefits of LSTMs over RNNs are less apparent, for this particular language modelling task.

On this note, trial 4 (RNN, subword tokenisation, dropout = 0.2) leads to the smallest testing loss and perplexity, displaying the strongest performance for this language modelling task.

* 1. **Qualitative Analysis (Text Generation)**

For each trial, text generation was performed with three different temperatures – 0.7, 1.0 and 1.3. “Today” was used as the starting word of choice. Here, we only show the generated text for trial 4 and trial 8, which are the best-performing trials for RNNs and LSTMs respectively. The generated text for all the trials can be found on the GitHub repository.

|  |  |  |
| --- | --- | --- |
| **Trial** | **Temperature** | **Generated Text (whitespace is represented as underscores (“\_”))** |
| 4 | 0.7 | Today ," ▁he ▁added . ▁" the ▁dollar 's ▁meeting ▁is ▁being ▁placed ▁on ▁the ▁outcome ▁of ▁collective ▁wage ▁agreement . ▁the ▁meeting ▁was ▁attended ▁by ▁farm ▁ministers ▁in ▁the ▁past ▁one ▁days ▁during ▁the ▁year - ago ▁week , ▁traders ▁said . ▁they ▁said ▁the ▁relaxation ▁of ▁controls ▁was ▁now . ▁sharp ▁details ▁were ▁not ▁disclosed . ▁the ▁company , ▁which ▁includes ▁a ▁69 ▁m ln ▁ dl rs ▁for ▁capital ▁spending , ▁less ▁than ▁half ▁of ▁the ▁amount ▁spent ▁70 . 6 ▁p ct ▁from ▁13 . 13 ▁m ln ▁in ▁ap ril , ▁and ▁brought ▁its ▁stake ▁in ▁s ime ▁ dar by , ▁a ▁spokesman ▁for ▁gr aan ▁elevator ▁m ij , ▁the ▁largest ▁employer ▁yesterday . ▁earlier ▁today , ▁the ▁company ▁said ▁the ▁transaction ▁involves ▁the ▁combination ▁of ▁oil ▁and ▁by ▁an ▁existing ▁20 ▁p ct ▁gain ▁in ▁w es ts ▁ uc cel ▁corp ▁as ▁pen ney 's ▁co ▁and ▁& lt ; format ura ▁ ini ez ione ▁po lim eri ▁spa >, ▁gen oa ▁and ▁& lt ; l om > ▁diesel ▁systems ▁and ▁other ▁company , ▁r oland ▁has ▁said ▁it ▁will ▁make ▁their ▁part ▁but ▁the ▁group ▁of ▁seven ▁-- ▁the ▁united ▁states , ▁would ▁be ▁the ▁critical ▁factor . ▁they ▁said ▁the ▁relaxation ▁of ▁controls ▁was ▁now ▁as ▁a ▁result ▁of ▁abroad ," ▁he ▁added . ▁ ▁" sq ui bb ▁is ▁putting ▁them , ▁it ▁said . ▁sci - med ▁year ley ▁inc ▁said ▁it ▁purchased ▁the ▁group ▁of ▁seven ▁ministers ▁in ▁the ▁gulf ▁to ▁country , ▁with ▁a ▁slightly ▁high ▁priority ▁from ▁the ▁price ▁of ▁the ▁united ▁states ▁with ▁reporting ▁such ▁as ▁prices ▁for ▁the |
| 1.0 | Today ▁vul ▁prime ▁policy ▁will ▁now ▁be ▁followed ▁by ▁the ▁potential ," ▁said gil lette . ▁shore ▁mentioned ▁half ▁its ▁ch il uba , ▁leader ▁of ▁the ▁z am bian ▁congress ▁in ▁the ▁u . s . ▁carries ▁out ▁a ▁threat ▁to ▁make ▁an ▁effort ▁trade ▁in ▁a ▁near - term ▁strike ▁on ▁t enders , ▁he ▁added ▁in ▁t ok yo ▁were ▁on ▁notice ▁this ▁week ▁because ▁of ▁the ▁j apan ese ▁economy ▁have ▁had ▁been ▁taken ▁higher ▁than ▁the ▁commercial ▁customers . ▁lower ▁taxes ▁would ▁be ▁early ▁by ▁the ▁u . s . ▁currency ▁proposal , ▁which ▁is ▁required ▁to ▁report ▁the ▁ bah rain ▁s mel ter ▁9 . ▁p ea body ▁and ▁co . ▁" there ▁is ▁ongoing , ▁serious ▁thought ▁applied ▁to ▁dome ▁in ▁its ▁retail ▁second ▁offer , ▁thus ▁ass uring ▁said ▁total ▁of ▁118 . 2 ▁m ln ▁s tg ▁the ▁j apan ese ▁early ▁income ▁goods ▁necessary , ▁and ▁that ▁it ▁would ▁be ▁taken ▁with ▁underlying ▁government ▁to ▁keep ▁our ▁trade ▁laws ." ▁" this ▁is ▁a ▁two - for - five ▁bonus ▁issue , ▁a ▁former ▁ argo sy st ems . ▁the ▁new ▁firm , ▁said ▁by ▁the ▁move ▁to ▁a ▁further ▁dollar ▁down ▁because ▁that ▁it ▁was ▁reached ▁in ▁new ▁y ork ▁and ▁seven , 000 ▁tonnes ▁of ▁wheat ▁for ▁an ▁e cu ▁gain ▁and ▁earlier ▁in ▁march ▁of ▁this ▁year 's ▁current ▁fiscal ▁quarter . ▁the ▁public ▁1985 / 87 ▁orders ▁at ▁end . 70 ▁m ln ▁tonnes , ▁vs ▁24 . 7 ▁m ln ▁six ▁months ▁shr ▁profit ▁eight ▁c ts ▁vs ▁loss ▁two ▁c ts ▁net ▁loss ▁ 725 , 000 ▁vs ▁profit ▁310 , |
| 1.3 | Today ▁also ▁will ▁not ▁do ▁anything ▁for ▁the ▁metal ▁rates . ▁" such ▁a ▁gathering ▁of ▁money ▁and ▁foreign ▁exchange ▁last ▁fixed ▁over ▁& lt ; bank burg ▁pin cus ▁and ▁co ▁& lt ; up cm . o > ▁and ▁a ▁shareholders ▁by ▁the ▁j apan ▁will ▁maintain ▁its ▁prime ▁and ▁both ▁electronic ▁abolish ▁foreign ▁investment ▁in ▁the ▁second ▁six ▁months ▁of ▁1987 . ▁loans ▁to ▁j une ▁17 ▁and ▁112 , 000 ▁b pd , ▁u . s ▁9 ▁and ▁2 . 99 ▁billion ▁ dl rs ▁from ▁86 ▁m ln ▁d lr ▁from ▁one ▁third ▁quarter ▁ended ▁ap ril ▁30 ▁shr ▁two . 2 ▁cents ▁with ▁the ▁rice ▁for ▁the ▁energy ▁cocoa ▁in ▁fiscal ▁leading ▁producer ▁are ▁need ▁to ▁put ▁who ▁have ▁some ▁leading ▁industrial ▁house ▁and ▁west ▁ger many 's ▁largely ▁" ▁goal ▁on ▁t uesday 's ▁west ▁g erman ▁monetary ▁policy ▁to ▁aid ▁the ▁dollar - if ▁end ▁and ▁there ▁are ▁are ▁likely ▁at ▁th over ▁banks . ▁" compan ies ▁like ▁dup ont ▁& lt ; dd > ▁ nova ▁said ▁the hor ▁offer ▁were ▁400 , 000 ▁tonnes ▁of ▁wheat ▁from ▁14 . 7 ▁m ln ▁u . s . ▁oil ▁as ▁under ▁information ▁new ▁y ork ▁denied ▁the ▁motion ▁here . ▁analysts ▁said ▁the ▁loss ▁of ▁the ▁inst itutes ▁will ▁become ▁improving ▁are ▁needed ▁with ▁stake ▁to ▁prevent ▁share ▁in ▁its ▁loan ▁negotiations , ▁17 . 0 ▁m ln ▁sw iss ▁f on 's ▁amount ▁to ▁end ▁their ▁s aj ▁paid . ▁ ira qi ▁troops ▁investors ▁in ▁tex as ▁mine ▁ak ▁co k ' ▁has ▁approved ▁an ▁agreement ▁for ▁12 ▁m ln ▁trader ▁bid ▁closes ▁on ▁behalf ▁of |
| 8 | 0.7 | Today ," ▁said ▁a ▁wall ▁street ▁arbit rag eur . ▁but ▁he ▁said ▁the ▁bank ▁regards ▁the ▁over draft ▁reference ▁rate , ▁based ▁on ▁short - term ▁rate ▁trends , ▁as ▁its ▁key ▁prime ▁lending ▁rate ▁to ▁corporate ▁customers . ▁the ▁loan ▁reference ▁rate ▁is ▁based ▁on ▁longer ▁term ▁trends . ▁the ▁bank ▁is ▁the ▁latest ▁to ▁cut ▁prime ▁rates ▁in ▁the ▁next ▁few ▁weeks . ▁given ▁an ▁average ▁yield ▁of ▁1 . 09 ▁billion ▁ dl rs ▁of ▁j uly ▁accounting ▁to ▁three ▁m ln ▁ dl rs ▁to ▁226 . 4 ▁m ln ▁ dl rs , ▁compared ▁with ▁169 . 2 ▁m ln ▁ dl rs ▁in ▁year ▁to ▁de cember ▁31 , ▁1986 . ▁shr ▁18 . 9 ▁m ln ▁vs ▁5 . 7 ▁m ln ▁note ▁- ▁full ▁name ▁is ▁data ▁services , ▁inc ▁said ▁it ▁will ▁offer ▁5 . 0 ▁m ln ▁ dl rs ▁cash ▁of ▁2 . 2 ▁m ln ▁common ▁shares ▁of ▁stock ▁for ▁each ▁ uc cel ▁wholly ▁based ▁on ▁50 ▁p ct ▁of ▁the ▁equity ▁in ▁j une , ▁it ▁added . ▁the ▁application ▁of ▁the ▁additional ▁48 ▁p ct ▁since ▁the ▁total ▁outstanding , ▁which ▁operates ▁46 ▁branches , ▁has ▁been ▁completed ▁in ▁principle ▁to ▁purchase ▁about ▁10 ▁m ln ▁va ▁shares ▁to ▁9 . 6 ▁m ln ▁ dl rs ▁of ▁fox ▁assets ▁of ▁n orc ros ▁p lc ▁& lt ; dl rs > ▁having ▁t bs , ▁which ▁is ▁worth ▁ab outh ▁55 ▁ dl rs ▁per ▁share . ▁" we 've ▁taken ▁up ▁with ▁all ▁alternatives ▁which ▁were ▁largely ▁complementary ▁unless ▁the ▁government 's ▁programme 's ▁current ▁management ▁and ▁tobacco ▁spirits . ▁ ▁the |
| 1.0 | Today >. ▁j apan 's ▁letter ▁follows ▁a ▁12 ▁p ct ▁increase ▁in ▁s ept ember ▁single - family ▁unit ▁starts ▁a ▁1 . 5 ▁p ct ▁to ▁1 . 73 ▁m ln ▁s tg , ▁it ▁is ▁also ▁most ▁for ▁good ▁recovery , ▁c omin co ▁said ▁in ▁view ▁that ▁levels ▁should ▁be ▁at ▁least ▁behind ▁this ▁time . ▁the ▁officials ▁said ▁ger many ▁had ▁practically ▁no ▁growth ▁in ▁the ▁long ▁term ▁and ▁bond ▁prices ▁was ▁not ▁immediately ▁clear ▁the ▁company ▁will ▁be ▁able ▁to ▁offer ▁specifics ▁from ▁the ▁sale ▁of ▁a ▁controlling ▁of ▁its ▁reliance ▁standard ▁life ▁insurance ▁co . ▁a ▁d alian ▁subsidiary , ▁r - held ▁dat eline ron ▁equipment ▁to ▁carry lift ▁as ▁s ears ▁ roe buck ▁and ▁co ▁& lt ; s > ▁converted ▁subsidiary ▁senior ▁management ▁of ▁international ▁data ▁corp ▁said ▁it ▁has ▁agreed ▁to ▁combine ▁its ▁cocoa ▁processing ▁businesses ▁with ▁those ▁of ▁s . ▁ . c . ▁terms , ▁effective ▁yesterday , ▁within ▁400 ▁miles ▁east ▁of ▁v ancouver , ▁produced ▁2 . 9 ▁m ln ▁barrels ▁a ▁year ▁earlier . ▁pre tax ▁earnings ▁fell ▁7 . 2 ▁p ct ▁to ▁3 . 48 ▁billion ▁pes os ▁to ▁the ▁dollar / 92 ▁to ▁17 . 6 ▁p ct ▁from ▁18 . 25 , ▁effective ▁tomorrow . ▁the ▁new ▁rate ▁is ▁based ▁on ▁a ▁basket ▁year ▁of ▁45 . 5 ▁p ct ▁this ▁year , ▁with ▁one ▁to ▁eight ▁billion ▁ dl rs ▁by ▁about ▁five ▁p ct . ▁in ▁an ▁initial ▁comment , ▁including ▁a ▁employee ▁stock ▁at ▁the ▁same ▁period ▁from ▁a ▁deficit ▁of ▁1 , 450 ▁billion ▁in ▁the ▁week ▁ended ▁ap ril ▁4 , ▁1987 , ▁sales ▁sales |
| 1.3 | Today 's ▁further ▁cut ▁since ▁they ▁should ▁earn ▁more ▁cash ▁or ▁before ▁in ▁fiscal ▁production ▁came . ▁speaking ▁at ▁a ▁forum ▁for ▁indust ria lists ▁of ▁major ▁central ▁banks ▁meeting ▁today 's ▁federal ▁reserve . ▁the ▁news ▁eroded ▁the ▁most ▁immediate ▁to ▁over ▁the ▁r eval uation ▁general ▁meeting , ▁the ▁meat ▁investment ▁firm ▁said . ▁sci - med ▁said ▁it ▁continues ▁to ▁be ▁identified ▁affected ▁by ▁ ron ald ▁per elman , ▁offered ▁20 ▁ dl rs ▁lower ▁than ▁an ▁average ▁yield ▁1 . 04 ▁ dl rs ▁in ▁liqu idation ▁value ▁of ▁a ▁stock ▁in ▁each ▁share ▁of ▁arrays ▁bought ▁or ▁less ▁than ▁2 . 4 ▁in ▁response ▁to ▁its ▁oil ▁market ." ▁ donald ▁trump , ▁rose ▁to ▁1 . 0 ▁m ln ▁ dl rs ▁reflecting ▁proposed ▁to ▁hol stein / 1 , 000 ▁when ▁h ilton ▁petroleum ▁spending ▁mill ▁workers ▁at ▁one ▁of ▁its ▁products , ▁at las ▁group , ▁prices ▁collapsed ▁by ▁the ▁end ▁of ▁an ▁oil ▁industry . ▁however , ▁government ▁officials ▁said . ▁au str alia 's ▁latest ▁rates ▁allot ▁are ▁not ▁going ▁to ▁be ▁more ▁early ." ▁bun des bank ▁officials ▁had ▁shown ▁fed ▁the ▁chances ▁to ▁be ▁what ▁i ran ▁know ▁before ▁the ▁fed ▁takes ▁sales ▁of ▁policy ▁than ▁a ▁15 ▁p ct ▁increase ▁in ▁gross ▁national ▁center ▁of ▁next ▁oct ober ▁as ▁of ▁export ▁subsidies ▁from ▁other ▁leading ▁paper ▁products , ▁the ▁gulf ▁for ▁ ku wait ▁tot alled ▁about ▁20 ▁billion ▁ dl rs ▁by ▁be cor 's ▁forecast ▁to ▁oct ▁1 . ▁he ▁noted ▁that ▁while ▁the ▁o pec ▁meeting ▁in ▁l j ens ▁to ▁buy ▁a ▁six ▁p ct ▁rate ▁rate ▁to ▁7 - 1 / |

The generated text is rather illogical, with the presence of frequent syntactic and grammatical errors. Semantic meaning can be inferred from very short sequences of words. However, limited continuity is observed beyond that. As a whole, the text itself generally does not convey a clear message, and there is little relationship between sentences. We can look at trial 4 (temperature = 1.3) as an example. The sequence “analysts said the loss of the institutes will become” makes sense, yet there is no mention of this “loss of the institutes” prior to it being mentioned. Moreover, what follows is oddly phrased and less reasonable (“improving are needed with stake to prevent share in its loan negotiations”).

As temperature increases, the text generated becomes more varied and less repetitive. In trial 4, a temperature value of 0.7 sees the same sequence “the relaxation of controls was now” being generated twice, while “the group of seven” also appears twice. However, a higher temperature value of 1.3 corresponds to greater creativity, with a wider range of tokens being used. Yet, this volatility also leads to the generation of more incomprehensible text, in which seemingly unrelated entities show up within the same sequence of tokens (eg. “iraqi troops investors in texas mine”).

Unlike the quantitative analysis, LSTMs arguably generate text of a slightly higher quality than RNNs – though they are still somewhat comparable. For a temperature value of 1.0, the RNN-generated text is largely devoid of meaning. There are very few instances of correctly phrased sequences – a notable example being “this year’s current fiscal quarter”. However, the LSTM-generated text has more occurrences of valid sequences. These sequences are also longer than those generated by the RNN. Examples include “the officials said germany had practically no growth in the long term”, “japan’s letter follows a 12 pct increase in” and “pretax earnings fell 7.2 pct to 3.48 billion pesos to the dollar”.

1. **Appendix**

I used GPT-5 to assist in the creation of code and improve the phrasing of the report. I am responsible for the content and quality of the submitted work. The GitHub repository for this assignment can be found at <https://github.com/chiabingxuan/Small-Language-Models-Comparison>.