NLP M2 REPORT
QA Model with context

A neural network for extracting or predicting answers of certain questions given the corresponding context paragraph.

Pre-processing

Pre-processing for Bi-lstm models

1. Loading the Dataset: The dataset used was the Stanford Question Answering Dataset (SQuAD).
I loaded the first 11,000 examples, keeping:

> **context**: paragraph from which the answer must be extracted.
> **question**: the question asked.
> **answers**: a dictionary containing the answer text and its character-level start index in the contex

2. Filtering & Shortening Contexts: To reduce outlier impact, I sorted examples by context length and selected the shortest 11,000 samples.

This helps the model train faster and prevents extreme padding.

3. Tokenization: I used Keras' Tokenizer to tokenize both context and question:
Fitted the tokenizer on all context and question texts combined to build a shared vocabulary.
Limited vocabulary size was computed as: vocab_size = len(tokenizer.word_index) + 1.

4. Padding Sequences: After converting words to sequences of indices:

Contexts were padded to *max_context_len* (e.g., 300 tokens).
Questions were padded to *max_question_len* (e.g., 30 tokens).
Padding was applied using pad_sequences(..., padding='post', truncating='post').

5. Aligning Answer Spans: The raw answers in SQuAD are character-based, while tokenized data is word-based.
I mapped each answer's character start and end positions to token indices using a custom function:

Iterated through token spans in context.
Matched answer boundaries to token positions.
This yielded start_token and end_token per sample.

6. One-Hot Encoding the Labels: The start and end positions were converted into one-hot vectors (length = max_context_len)

7. Train-Test Split:

Final dataset was split into:

Training set: first 10,000 samples.
Test set: next 1,000 samples.

Input shapes:

padded_contexts: (samples, max_context_len)
padded_questions: (samples, max_question_len)
y_start, y_end: one-hot encoded vectors

Pre-processing for transformers

1. Tokenization

Used a pretrained BERT tokenizer to tokenize the question and context together.
The tokenizer generated:
Token IDs.
Attention masks.
Offset mappings (linking tokens to their original character positions in the context).

2. Answer Span Alignment

Since SQuAD provides the answer's character-level start index:
The offset mappings were used to convert this into token-level start and end positions.
These token indices serve as the model's target outputs (labels) for start and end positions.

3. Input Standardization

All sequences (inputs and masks) were padded or truncated to a fixed maximum length
This ensured uniform shape across all samples, allowing efficient batch processing.

4. Output Label Creation

For each sample, two labels were created:

One-hot position of the **start token**,

One-hot position of the **end token**,
Based on the aligned token indices of the answer span.

5. Dataset Preparation

The processed inputs (token IDs and attention masks) and labels (start/end positions) were grouped into a training dataset.
The dataset was:

Shuffled to prevent bias from input order,
Batched to enable parallel processing,
Prefetched to reduce training latency and improve performance.

Post-processing

Post-processing for the BI-LSTM Model

1. Token Probability Interpretation:
For each test sample, the model output two probability distributions:

One for the **start token** position.
One for the **end token** position.

Instead of taking a simple argmax, a custom **best span extraction algorithm** was used to:

Ensure that the end token comes **after** the start token.
Limit the **answer span length** (e.g. ≤ 15 tokens).
Select the best start–end pair that **maximizes the joint probability** (start × end).

2. Token-to-Text Mapping:
The contexts were tokenized using Keras' text_to_word_sequence, which breaks the context into a clean list of words (tokens).
Once the predicted token indices (start_idx, end_idx) were obtained, the corresponding **text span** was reconstructed using:

tokens[start_idx:end_idx + 1]
This slice gave the exact group of words predicted as the answer.
The final **predicted answer** was formed by joining these tokens back into a string.

3. Comparing to True Answer:
The true start and end labels (from preprocessing) were **one-hot encoded**, so we used argmax() over y_start and y_end to get the **true token indices**.
These true indices were then used to reconstruct the **ground-truth token-mapped answer** for comparison.

4. Raw Answer Reference:
To validate token-mapping accuracy, the original **character-based answer** was also retrieved directly from the raw context using the stored character offsets.

post-processing for the transformers

After the model produced its predictions (start and end logits), I applied a post-processing step to extract the final answer span from the raw context. The key steps involved were:

- Tokenizing the question and context again using the BERT tokenizer, while storing **offset mappings** that track the character positions of each token in the original context.
- Feeding the tokenized input into the model to obtain the **logits** for both start and end positions.
- Selecting the most probable start and end token indices using **argmax** over the predicted logits.
- Handling edge cases where the predicted end position was before the start, by adjusting the end index accordingly.
- Using the **offset mappings** to translate the predicted token indices back into character positions.
- Extracting and returning the **actual text span** from the context based on the predicted character range.

System architecture and pipeline

BI-LSTM:

1. BI-LSTM Model Architecture Overview

The question answering system is built using a **three-layer Bi-LSTM architecture**:

Inputs:
Two sequences: the **context** and the **question**, both tokenized and padded.

Embedding:
A shared embedding layer transforms token IDs into dense vectors.

Hidden Layers:

BiLSTM on Context: Captures bidirectional context for each word.
BiLSTM on Question: Produces a global question vector.
Fusion Layer: Combines context and question using concatenation, element-wise subtraction, and multiplication, followed by a dense layer with ReLU activation.

Outputs:
Two softmax layers predict the **start** and **end** token positions of the answer.

2. Training Pipeline Summary

Shuffling and Splitting:
Entire dataset is shuffled, then split into **training** and **validation** sets.

Training:

Optimizer: Adam
Loss: Categorical cross-entropy for start and end outputs.
Metrics: Accuracy for both start and end positions.

Transformer

1. Transformer-Based Model Architecture

This model is built using a custom Transformer encoder architecture designed for span-based question answering.

Input:
input_ids and attention_mask for each question-context pair, padded to a fixed length.

Embedding Layer:
Combines token embeddings and positional embeddings to preserve word identity and order.

Transformer Encoder Layers ***(2 layers used)***:
Each encoder includes:

Multi-head self-attention to model contextual dependencies.
Feedforward projection layer.
Residual connections, layer normalization, and dropout for regularization.

Output Heads:
Two separate dense layers produce start and end logits over the context tokens.

2. Training Pipeline Summary

Loss Function:
Uses SparseCategoricalCrossentropy on token-level predictions (start and end indices).

Compilation:
Optimizer: Adam with a learning rate of 1e-4.
Metrics: Accuracy for both start and end token predictions.

Training:
The model is trained on preprocessed TensorFlow Dataset batches.
Runs for 15 epochs, tracking loss and accuracy.

Training process

Choice of loss function

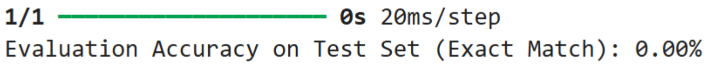
categorical_crossentropy is the right loss, because:

The output is a softmax probability distribution.
The labels are categorical (one correct index per sequence).
It penalizes the model based on the negative log-probability of the correct class.

Methodology: 7 Trials

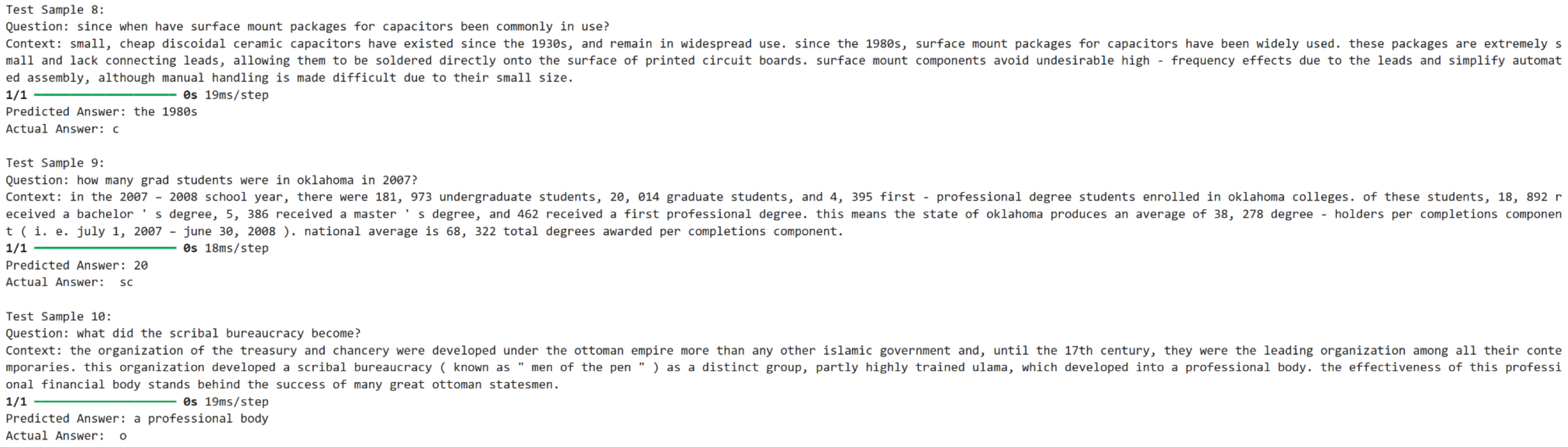
Transformers:

1. Extractor: *(testing on dummy data)*
Not predicting
Training accuracy: almost zero



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 20ms/step
Evaluation Accuracy on Test Set (Exact Match): 0.00%
```

2. Generator: *(testing on dummy data)*
Not predicting
Training accuracy: 70%



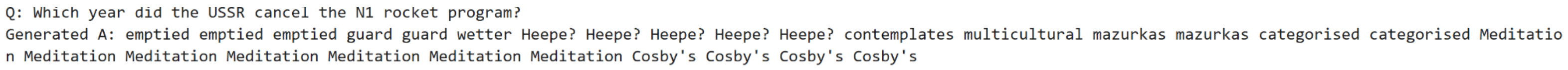
```
Test Sample 8:
Question: since when have surface mount packages for capacitors been commonly in use?
Context: small, cheap discoidal ceramic capacitors have existed since the 1930s, and remain in widespread use. since the 1980s, surface mount packages for capacitors have been widely used. these packages are extremely s
mall and lack connecting leads, allowing them to be soldered directly onto the surface of printed circuit boards. surface mount components avoid undesirable high - frequency effects due to the leads and simplify automat
ed assembly, although manual handling is made difficult due to their small size.
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 19ms/step
Predicted Answer: the 1980s
Actual Answer: c

Test Sample 9:
Question: how many grad students were in oklahoma in 2007?
Context: in the 2007 – 2008 school year, there were 181, 973 undergraduate students, 20, 014 graduate students, and 4, 395 first - professional degree students enrolled in oklahoma colleges. of these students, 18, 892 r
eceived a bachelor ' s degree, 5, 386 received a master ' s degree, and 462 received a first professional degree. this means the state of oklahoma produces an average of 38, 278 degree - holders per completions componen
t ( i. e. july 1, 2007 – june 30, 2008 ). national average is 68, 322 total degrees awarded per completions component.
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step
Predicted Answer: 20
Actual Answer:  sc

Test Sample 10:
Question: what did the scribal bureaucracy become?
Context: the organization of the treasury and chancery were developed under the ottoman empire more than any other islamic government and, until the 17th century, they were the leading organization among all their conte
mporaries. this organization developed a scribal bureaucracy ( known as " men of the pen " ) as a distinct group, partly highly trained ulama, which developed into a professional body. the effectiveness of this professi
onal financial body stands behind the success of many great ottoman statesmen.
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 19ms/step
Predicted Answer: a professional body
Actual Answer:  o
```

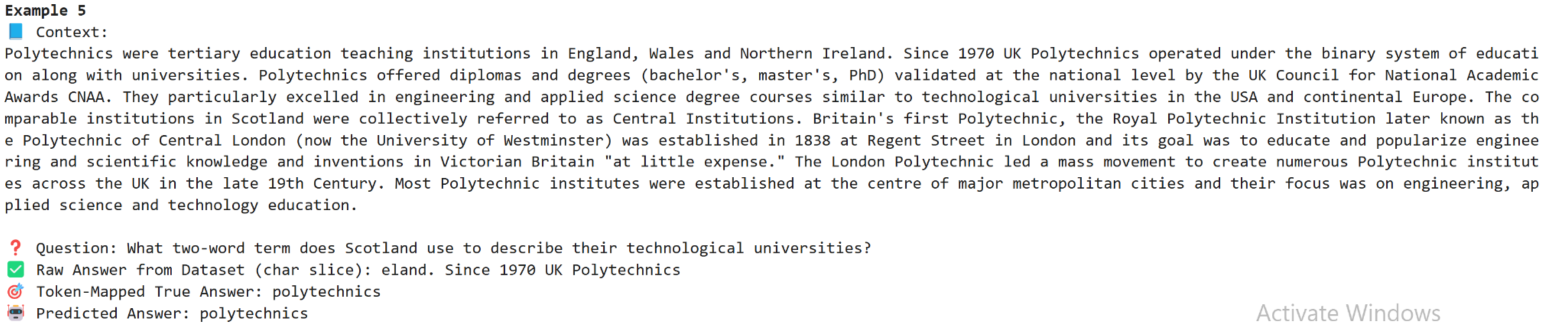
LSTMs:

*1. **seq2seq:** (testing on dummy data)*
Training accuracy: 90%



```
Q: Which year did the USSR cancel the N1 rocket program?
Generated A: emptied emptied emptied guard guard wetter Heepe? Heepe? Heepe? Heepe? Heepe? contemplates multicultural mazurkas mazurkas categorised categorised Meditatio
n Meditation Meditation Meditation Meditation Meditation Meditation Cosby's Cosby's Cosby's Cosby's
```

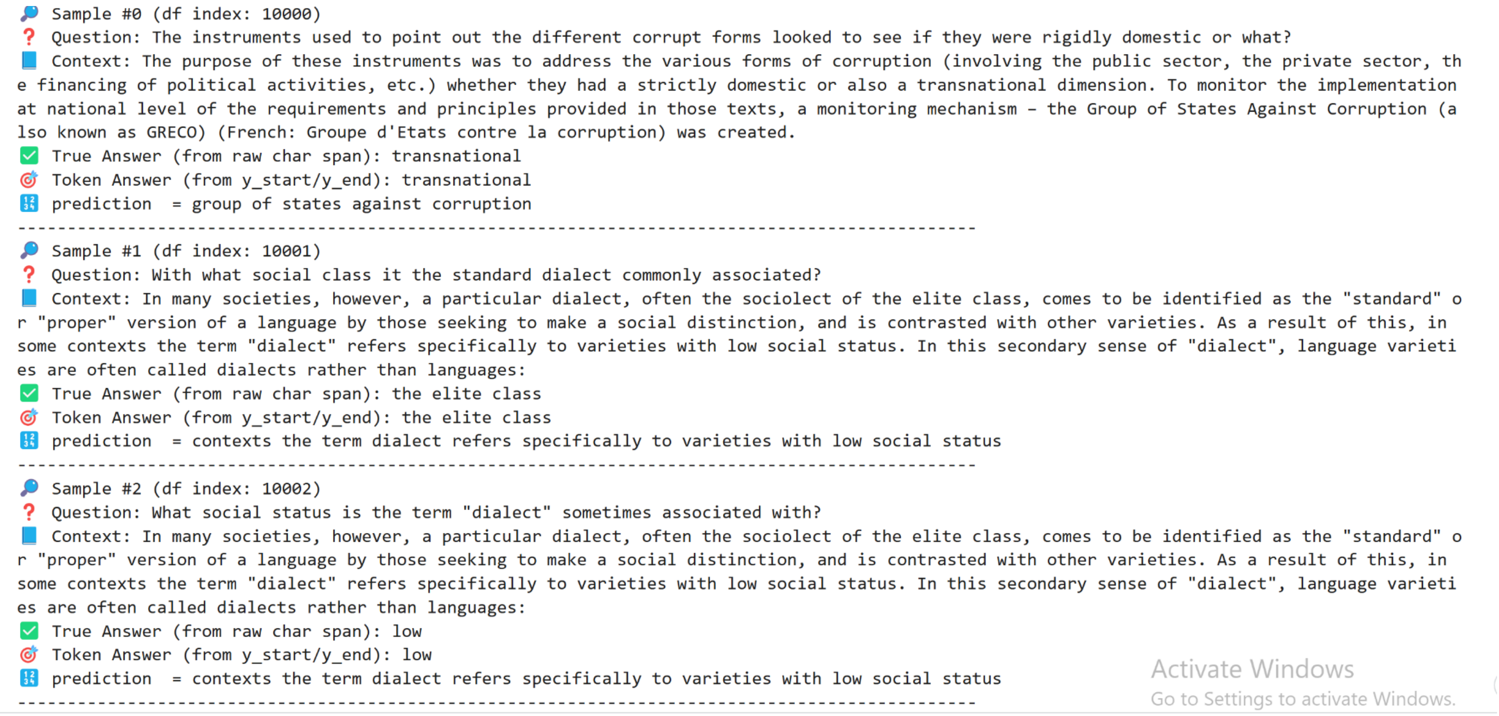
2. BI-LSTM Trial 1:
Training accuracy: 40%



```
Example 5
📘 Context:
Polytechnics were tertiary education teaching institutions in England, Wales and Northern Ireland. Since 1970 UK Polytechnics operated under the binary system of educati
on along with universities. Polytechnics offered diplomas and degrees (bachelor's, master's, PhD) validated at the national level by the UK Council for National Academic
Awards CNAA. They particularly excelled in engineering and applied science degree courses similar to technological universities in the USA and continental Europe. The co
mparable institutions in Scotland were collectively referred to as Central Institutions. Britain's first Polytechnic, the Royal Polytechnic Institution later known as th
e Polytechnic of Central London (now the University of Westminster) was established in 1838 at Regent Street in London and its goal was to educate and popularize enginee
ring and scientific knowledge and inventions in Victorian Britain "at little expense." The London Polytechnic led a mass movement to create numerous Polytechnic institut
es across the UK in the late 19th Century. Most Polytechnic institutes were established at the centre of major metropolitan cities and their focus was on engineering, ap
plied science and technology education.

❓ Question: What two-word term does Scotland use to describe their technological universities?
✅ Raw Answer from Dataset (char slice): eland. Since 1970 UK Polytechnics
🎯 Token-Mapped True Answer: polytechnics
🤖 Predicted Answer: polytechnics
```

3. BI-LSTM Trial 2:
Training accuracy: 90%



```
🔎 Sample #0 (df index: 10000)
❓ Question: The instruments used to point out the different corrupt forms looked to see if they were rigidly domestic or what?
📘 Context: The purpose of these instruments was to address the various forms of corruption (involving the public sector, the private sector, th
e financing of political activities, etc.) whether they had a strictly domestic or also a transnational dimension. To monitor the implementation
at national level of the requirements and principles provided in those texts, a monitoring mechanism – the Group of States Against Corruption (a
lso known as GRECO) (French: Groupe d'Etats contre la corruption) was created.
✅ True Answer (from raw char span): transnational
🎯 Token Answer (from y_start/y_end): transnational
🔢 prediction  = group of states against corruption
--------------------------------------------------------------------------------
🔎 Sample #1 (df index: 10001)
❓ Question: With what social class it the standard dialect commonly associated?
📘 Context: In many societies, however, a particular dialect, often the sociolect of the elite class, comes to be identified as the "standard" o
r "proper" version of a language by those seeking to make a social distinction, and is contrasted with other varieties. As a result of this, in
some contexts the term "dialect" refers specifically to varieties with low social status. In this secondary sense of "dialect", language varieti
es are often called dialects rather than languages:
✅ True Answer (from raw char span): the elite class
🎯 Token Answer (from y_start/y_end): the elite class
🔢 prediction  = contexts the term dialect refers specifically to varieties with low social status
--------------------------------------------------------------------------------
🔎 Sample #2 (df index: 10002)
❓ Question: What social status is the term "dialect" sometimes associated with?
📘 Context: In many societies, however, a particular dialect, often the sociolect of the elite class, comes to be identified as the "standard" o
r "proper" version of a language by those seeking to make a social distinction, and is contrasted with other varieties. As a result of this, in
some contexts the term "dialect" refers specifically to varieties with low social status. In this secondary sense of "dialect", language varieti
es are often called dialects rather than languages:
✅ True Answer (from raw char span): low
🎯 Token Answer (from y_start/y_end): low
🔢 prediction  = contexts the term dialect refers specifically to varieties with low social status
--------------------------------------------------------------------------------
```

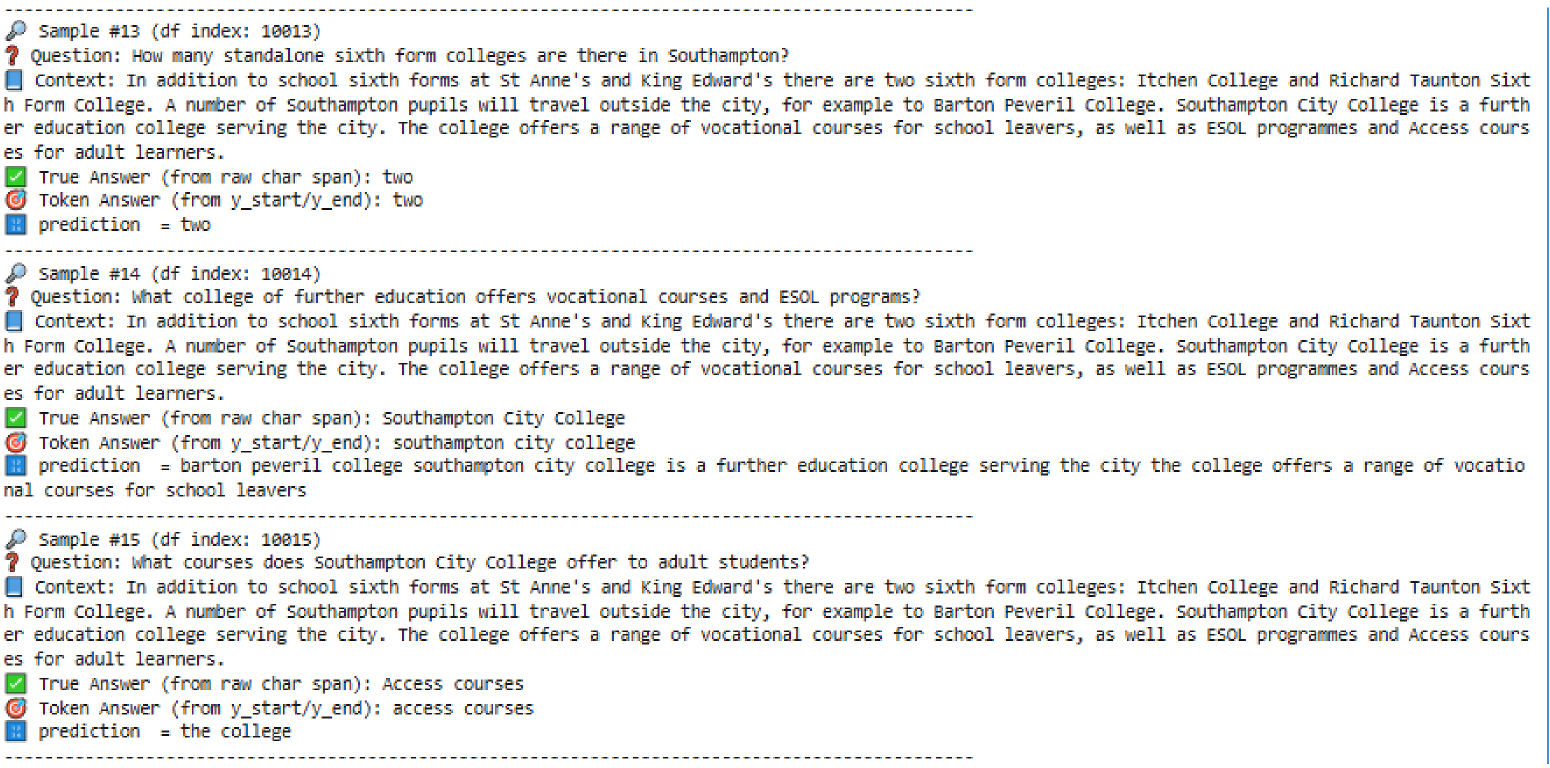
The Problem was: it's not question aware at all, it assigns same exact answer for the context, even if we shuffled the context order in the training, it still not paying attention to the question itself

Knowing that:

- It predicts answers that (contains) the right answer for one question out of the many ones on the same context
- It predicted the exact same answer for that same question, when the context was passed to the

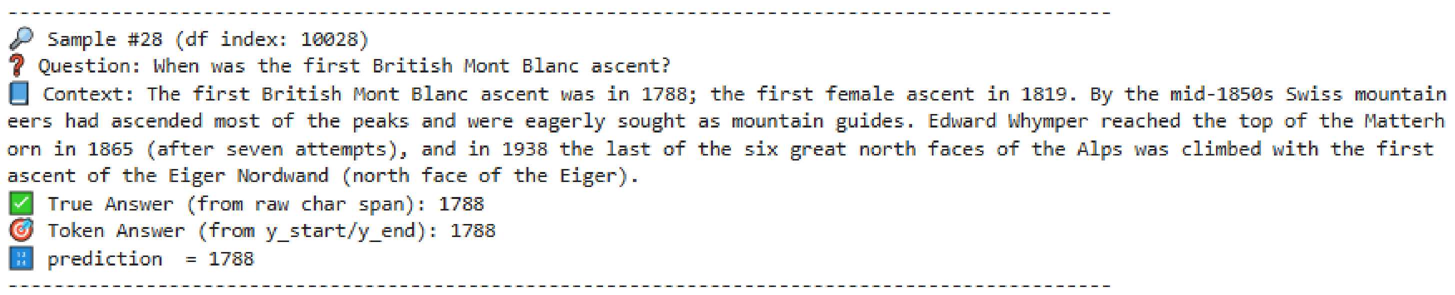
The following are trials after attempting to solve the problem of (non-question aware prediction)

4. BI-LSTM Trial 3: *(MULTIPLE LAYERS)*
Training accuracy: 92%

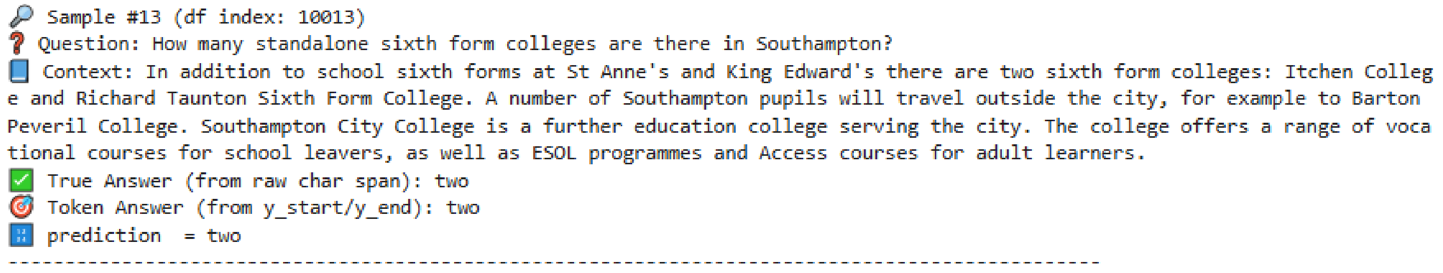


```
--------------------------------------------------------------------------------
🔎 Sample #13 (df index: 10013)
❓ Question: How many standalone sixth form colleges are there in Southampton?
📘 Context: In addition to school sixth forms at St Anne's and King Edward's there are two sixth form colleges: Itchen College and Richard Taunton Sixt
h Form College. A number of Southampton pupils will travel outside the city, for example to Barton Peveril College. Southampton City College is a furth
er education college serving the city. The college offers a range of vocational courses for school leavers, as well as ESOL programmes and Access cours
es for adult learners.
✅ True Answer (from raw char span): two
🎯 Token Answer (from y_start/y_end): two
🔢 prediction  = two
--------------------------------------------------------------------------------
🔎 Sample #14 (df index: 10014)
❓ Question: What college of further education offers vocational courses and ESOL programs?
📘 Context: In addition to school sixth forms at St Anne's and King Edward's there are two sixth form colleges: Itchen College and Richard Taunton Sixt
h Form College. A number of Southampton pupils will travel outside the city, for example to Barton Peveril College. Southampton City College is a furth
er education college serving the city. The college offers a range of vocational courses for school leavers, as well as ESOL programmes and Access cours
es for adult learners.
✅ True Answer (from raw char span): Southampton City College
🎯 Token Answer (from y_start/y_end): southampton city college
🔢 prediction  = barton peveril college southampton city college is a further education college serving the city the college offers a range of vocatio
nal courses for school leavers
--------------------------------------------------------------------------------
🔎 Sample #15 (df index: 10015)
❓ Question: What courses does Southampton City College offer to adult students?
📘 Context: In addition to school sixth forms at St Anne's and King Edward's there are two sixth form colleges: Itchen College and Richard Taunton Sixt
h Form College. A number of Southampton pupils will travel outside the city, for example to Barton Peveril College. Southampton City College is a furth
er education college serving the city. The college offers a range of vocational courses for school leavers, as well as ESOL programmes and Access cours
es for adult learners.
✅ True Answer (from raw char span): Access courses
🎯 Token Answer (from y_start/y_end): access courses
🔢 prediction  = the college
--------------------------------------------------------------------------------
```

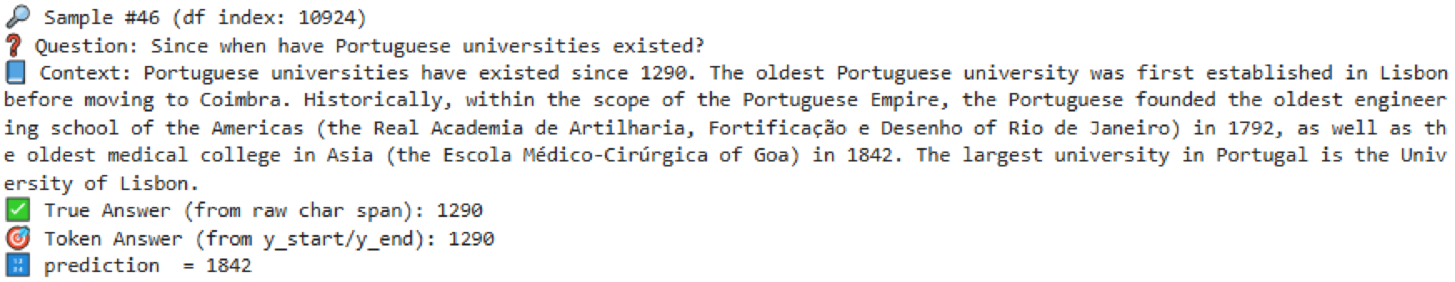
Here the model predicted 3 different answers for the same answer
1st one exactly matches the true answer
2nd one has the true answer within it
3rd one predicted wrong answer



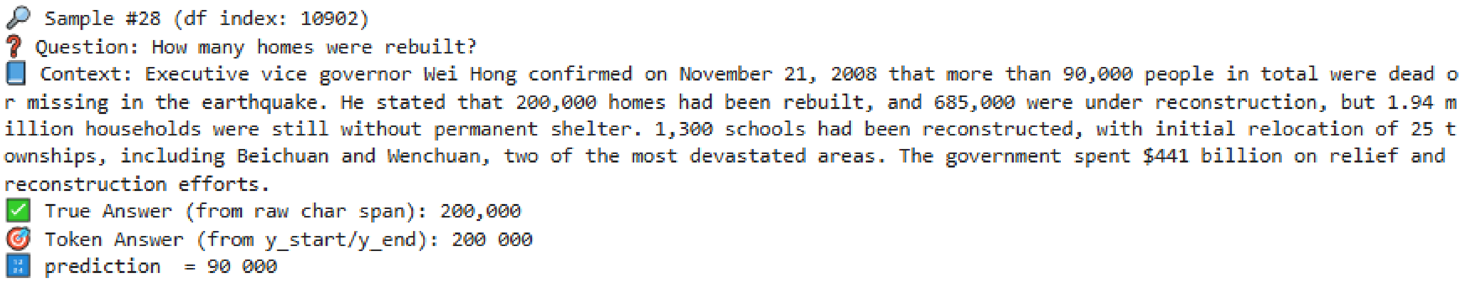
```
------------------------------------------------------------------------------------------
🔎 Sample #28 (df index: 10028)
❓ Question: When was the first British Mont Blanc ascent?
📘 Context: The first British Mont Blanc ascent was in 1788; the first female ascent in 1819. By the mid-1850s Swiss mountain
eers had ascended most of the peaks and were eagerly sought as mountain guides. Edward Whymper reached the top of the Matterh
orn in 1865 (after seven attempts), and in 1938 the last of the six great north faces of the Alps was climbed with the first
ascent of the Eiger Nordwand (north face of the Eiger).
✅ True Answer (from raw char span): 1788
🎯 Token Answer (from y_start/y_end): 1788
🔢 prediction  = 1788
------------------------------------------------------------------------------------------
```



```
🔎 Sample #13 (df index: 10013)
❓ Question: How many standalone sixth form colleges are there in Southampton?
📘 Context: In addition to school sixth forms at St Anne's and King Edward's there are two sixth form colleges: Itchen Colleg
e and Richard Taunton Sixth Form College. A number of Southampton pupils will travel outside the city, for example to Barton
Peveril College. Southampton City College is a further education college serving the city. The college offers a range of voca
tional courses for school leavers, as well as ESOL programmes and Access courses for adult learners.
✅ True Answer (from raw char span): two
🎯 Token Answer (from y_start/y_end): two
🔢 prediction  = two
------------------------------------------------------------------------------------------
```



```
🔎 Sample #46 (df index: 10924)
❓ Question: Since when have Portuguese universities existed?
📘 Context: Portuguese universities have existed since 1290. The oldest Portuguese university was first established in Lisbon
before moving to Coimbra. Historically, within the scope of the Portuguese Empire, the Portuguese founded the oldest engineer
ing school of the Americas (the Real Academia de Artilharia, Fortificação e Desenho of Rio de Janeiro) in 1792, as well as th
e oldest medical college in Asia (the Escola Médico-Cirúrgica of Goa) in 1842. The largest university in Portugal is the Univ
ersity of Lisbon.
✅ True Answer (from raw char span): 1290
🎯 Token Answer (from y_start/y_end): 1290
🔢 prediction  = 1842
```

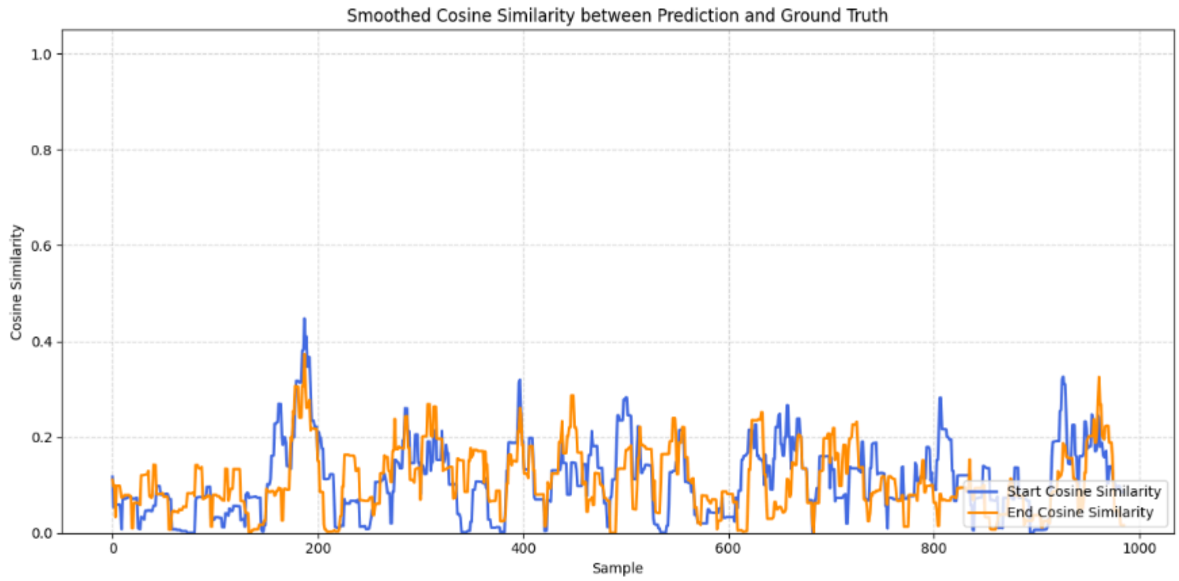


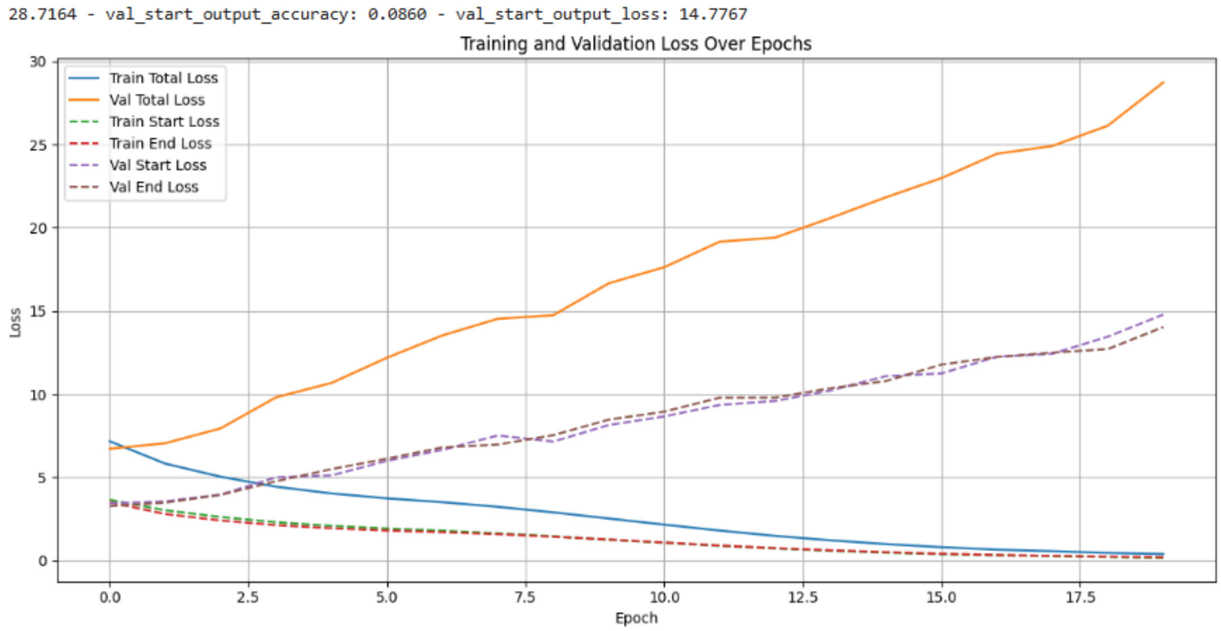
```
🔎 Sample #28 (df index: 10902)
❓ Question: How many homes were rebuilt?
📘 Context: Executive vice governor Wei Hong confirmed on November 21, 2008 that more than 90,000 people in total were dead o
r missing in the earthquake. He stated that 200,000 homes had been rebuilt, and 685,000 were under reconstruction, but 1.94 m
illion households were still without permanent shelter. 1,300 schools had been reconstructed, with initial relocation of 25 t
ownships, including Beichuan and Wenchuan, two of the most devastated areas. The government spent $441 billion on relief and
reconstruction efforts.
✅ True Answer (from raw char span): 200,000
🎯 Token Answer (from y_start/y_end): 200 000
🔢 prediction  = 90 000
```

In some examples it learned that how many is associated with numbers and when is associated with dates so it predicted numbers but not the correct one each time

Cosine similarity and loss plot

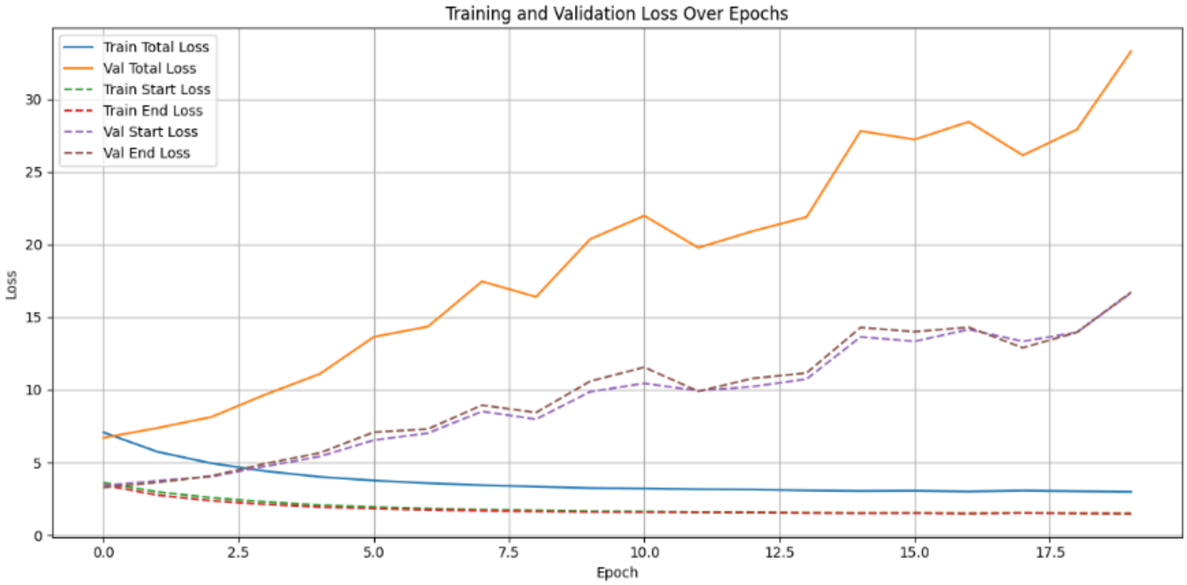
We started to plot the cosine similarity and loss for the trials we seemed satisfied with

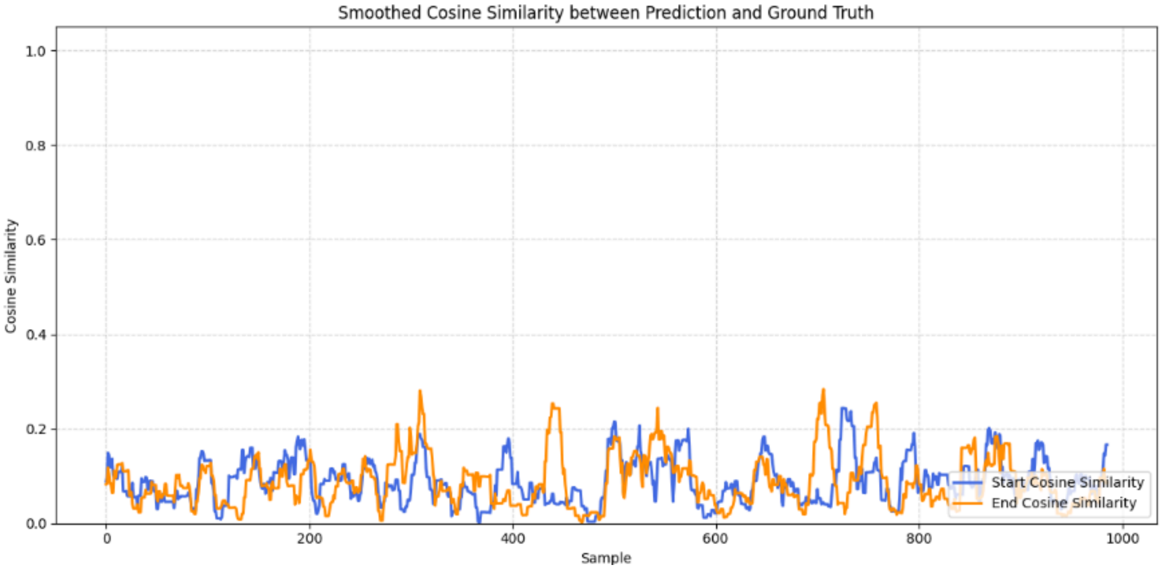


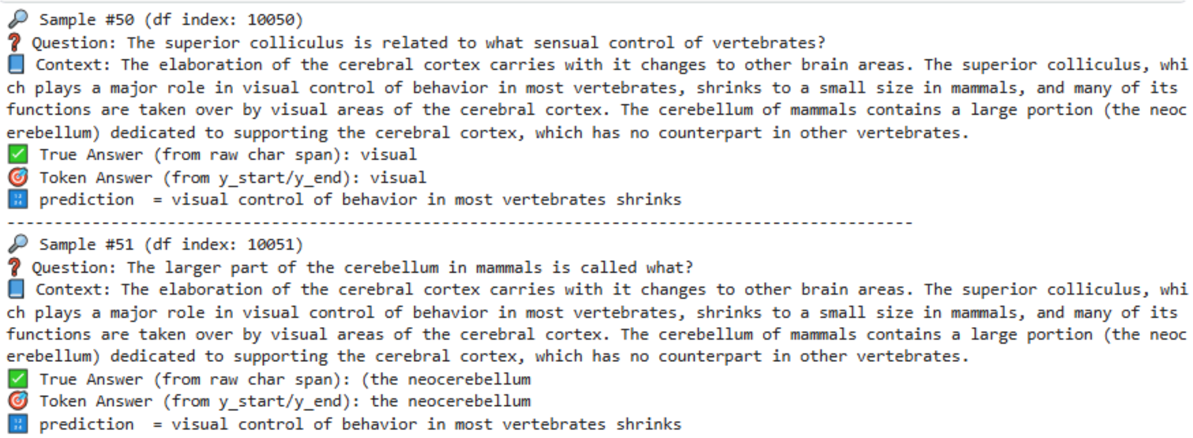


```
28.7164 - val_start_output_accuracy: 0.0860 - val_start_output_loss: 14.7767
```

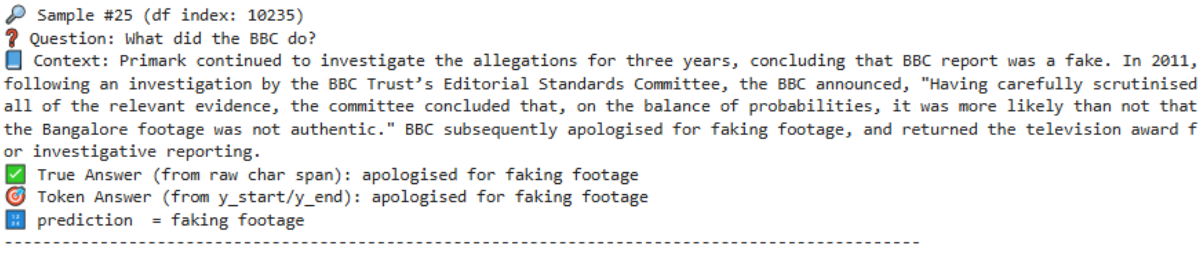
5. BI-LSTM Trial 4: (3 HIDDEN LAYERS ONLY)



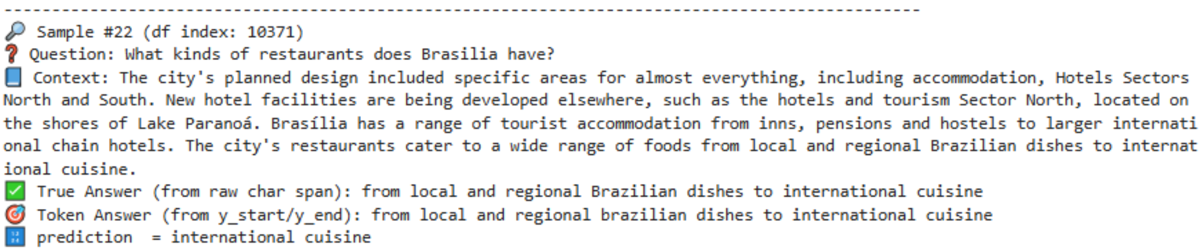




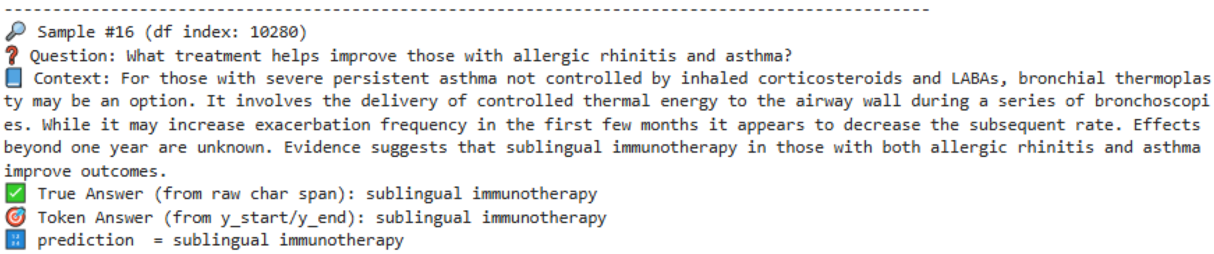
```
🔎 Sample #50 (df index: 10050)
❓ Question: The superior colliculus is related to what sensual control of vertebrates?
📘 Context: The elaboration of the cerebral cortex carries with it changes to other brain areas. The superior colliculus, whi
ch plays a major role in visual control of behavior in most vertebrates, shrinks to a small size in mammals, and many of its
functions are taken over by visual areas of the cerebral cortex. The cerebellum of mammals contains a large portion (the neoc
erebellum) dedicated to supporting the cerebral cortex, which has no counterpart in other vertebrates.
✅ True Answer (from raw char span): visual
🎯 Token Answer (from y_start/y_end): visual
🔢 prediction  = visual control of behavior in most vertebrates shrinks
------------------------------------------------------------------------------------------
🔎 Sample #51 (df index: 10051)
❓ Question: The larger part of the cerebellum in mammals is called what?
📘 Context: The elaboration of the cerebral cortex carries with it changes to other brain areas. The superior colliculus, whi
ch plays a major role in visual control of behavior in most vertebrates, shrinks to a small size in mammals, and many of its
functions are taken over by visual areas of the cerebral cortex. The cerebellum of mammals contains a large portion (the neoc
erebellum) dedicated to supporting the cerebral cortex, which has no counterpart in other vertebrates.
✅ True Answer (from raw char span): (the neocerebellum
🎯 Token Answer (from y_start/y_end): the neocerebellum
🔢 prediction  = visual control of behavior in most vertebrates shrinks
```



```
🔎 Sample #25 (df index: 10235)
❓ Question: What did the BBC do?
📘 Context: Primark continued to investigate the allegations for three years, concluding that BBC report was a fake. In 2011,
following an investigation by the BBC Trust’s Editorial Standards Committee, the BBC announced, "Having carefully scrutinised
all of the relevant evidence, the committee concluded that, on the balance of probabilities, it was more likely than not that
the Bangalore footage was not authentic." BBC subsequently apologised for faking footage, and returned the television award f
or investigative reporting.
✅ True Answer (from raw char span): apologised for faking footage
🎯 Token Answer (from y_start/y_end): apologised for faking footage
🔢 prediction  = faking footage
------------------------------------------------------------------------------------------
```



```
------------------------------------------------------------------------------------------
🔎 Sample #22 (df index: 10371)
❓ Question: What kinds of restaurants does Brasilia have?
📘 Context: The city's planned design included specific areas for almost everything, including accommodation, Hotels Sectors
North and South. New hotel facilities are being developed elsewhere, such as the hotels and tourism Sector North, located on
the shores of Lake Paranoá. Brasília has a range of tourist accommodation from inns, pensions and hostels to larger internati
onal chain hotels. The city's restaurants cater to a wide range of foods from local and regional Brazilian dishes to internat
ional cuisine.
✅ True Answer (from raw char span): from local and regional Brazilian dishes to international cuisine
🎯 Token Answer (from y_start/y_end): from local and regional brazilian dishes to international cuisine
🔢 prediction  = international cuisine
```



```
------------------------------------------------------------------------------------------
🔎 Sample #16 (df index: 10280)
❓ Question: What treatment helps improve those with allergic rhinitis and asthma?
📘 Context: For those with severe persistent asthma not controlled by inhaled corticosteroids and LABAs, bronchial thermoplas
ty may be an option. It involves the delivery of controlled thermal energy to the airway wall during a series of bronchoscopi
es. While it may increase exacerbation frequency in the first few months it appears to decrease the subsequent rate. Effects
beyond one year are unknown. Evidence suggests that sublingual immunotherapy in those with both allergic rhinitis and asthma
improve outcomes.
✅ True Answer (from raw char span): sublingual immunotherapy
🎯 Token Answer (from y_start/y_end): sublingual immunotherapy
🔢 prediction  = sublingual immunotherapy
```

Limitations and Suggested Improvements

- custom loss- not guaranteed.

- Layers and Units: A model that's too shallow or with too few units might not be able to learn enough from the data. Experiment with adding more layers or increasing the hidden units.

- **Generative task**: One notable limitation encountered during this project was the attempted use of a sequence-to-sequence (seq2seq) architecture, which ultimately proved to be unsuitable for the task. Unlike tasks such as machine translation, where the model must generate novel sequences, the question answering task addressed here is extractive in nature—meaning the answer is always a span within the given context, not something that needs to be generated. As a result, generative architectures like seq2seq did not align well with the task requirements.

- **Integrating mechanisms**: Additionally, many state-of-the-art models for extractive question answering commonly integrate both attention mechanisms and recurrent structures (e.g., Bi-LSTMs) to enhance performance. However, due to the project constraints, we were restricted from combining both components within the same architecture. This limited our ability to fully leverage the synergistic benefits of hybrid attention-based RNN models that have proven effective in prior research.