1. Which of the following are bottlenecks when implementing seq2seq models? 1 / 1 point

- [x] You are trying to store variable length sequences in a fixed memory, for example, you are trying to store articles of different lengths in a fixed 100 dimensional vector.

  Correct
  Correct

- [x] There are vanishing/exploding gradient problems.

  Correct
  Correct

- [ ] They require a lot of memory.
- [ ] They can directly handle sequences of variable lengths without any modification.

2. What are some of the benefits of using attention? 1 / 1 point

- [x] It improves the interpretability of the model by highlighting which parts of the input contribute to the output.

  Correct
  Correct

- [x] It allows you to focus on the parts that matter more.

  Correct
  Correct.

- [x] It helps with the information bottleneck issue.

  Correct
  Correct.

- [x] It simplifies the model architecture by reducing the need for complex recurrent layers.

  Correct
  Correct.

3. In the context of Transformer models, which components are essential for the attention mechanism? Select all that apply. 1 / 1 point

- [x] **Keys:** Described as part of the mechanism that helps in identifying relevant information from the input data.

  Correct
  Keys help the model pick out the relevant bits from your data, matching them with queries.

- [ ] **Activation Functions:** While activation functions like ReLU are crucial in neural network layers for introducing non-linearity, they are not considered an essential component of the attention mechanism itself. (Incorrect) - This replaces the previously correct option.

- [x] **Queries:** Described as the component that represents the current item being processed, to find matching information.

  Correct
  Queries act like a spotlight, focusing on what's currently important in the model's task.

- [x] **Values:** These hold the actual information from the input that will be used to compute the output, once a match is found between a key and a query.

  Correct
  Once the model finds a match between keys and queries, values provide the actual information we want to use.

- [ ] **Cosine similarity:** While useful in measuring similarity between vectors, it is not an essential component of the attention mechanism in Transformer models.

4. Teacher forcing uses the actual output from the training dataset at time step $y^{(t)}$ as input in the next time step $X^{(t+1)}$, instead of the output generated by your model. 1 / 1 point

- (•) True.
- ( ) False.

Correct
Correct.

5. The BLEU score's range is as follows: 1 / 1 point

- (●) The closer to 0, the worse it is, the closer to 1, the better it is.
- ( ) The closer to 1, the worse it is, the closer to 0, the better it is.
- ( ) The closer to -1, the worse it is, the closer to 1, the better it is.
- ( ) The closer to $-\infty$, the worse it is, the closer to ∞, the better it is.

Correct
Correct.

6. BLEU (Vanilla Implementation) is defined as: 1 / 1 point

- (●) (Sum of unique n-gram counts, overlapping in the candidate and reference) / (Total # of n-grams in the candidate)
- ( ) (Sum of unique n-gram counts in the candidate) / (Total # of n-grams in the candidate)
- ( ) (Sum of unique unigram counts, overlapping in the candidate and reference) / (Total # of unigrams in the reference)
- ( ) (Sum of unique unigram counts in the candidate) / (Total # of n-grams in the reference)

Correct
Correct.

7. What aspect of text summaries does the Rouge metric primarily evaluate? 1 / 1 point

- ( ) The uniqueness of words in the summary.
- ( ) Rouge doesn't evaluate the summary based on its length. It's more about comparing the content to reference summaries.
- (●) The similarity of the summary to reference summaries.
- ( ) The grammatical correctness of the summary.

Correct
Rouge measures how well the summary captures the key points found in reference summaries. It's all about matching content.

- ○ Rouge doesn't evaluate the summary based on its length. It's more about comparing the content to reference summaries.
- ◉ The similarity of the summary to reference summaries.
- ○ The grammatical correctness of the summary.

Correct
Rouge measures how well the summary captures the key points found in reference summaries. It's all about matching content.

8. Greedy decoding — 1 / 1 point

- ◉ Allows you select the word with the highest probability at each time step.
- ○ Allows you randomly select the word according to its own probability in the softmax layer.
- ○ Selects multiple options for the best input based on conditional probability.
- ○ Makes use of the Minimum Bayes Risk method.

Correct
Correct.

9. When implementing Minimum Bayes Risk method in decoding, let's say with 4 samples, you have to implement the following. — 1 / 1 point

1. Calculate similarity score between sample 1 and sample 2
2. Calculate similarity score between sample 1 and sample 3
3. Calculate similarity score between sample 1 and sample 4
4. Average the score of the first 3 steps (Usually a weighted average)
5. Repeat until all samples have overall scores

Pick the best candidate, with the highest similarity score.

- ◉ True
- ○ False

Correct
Correct.