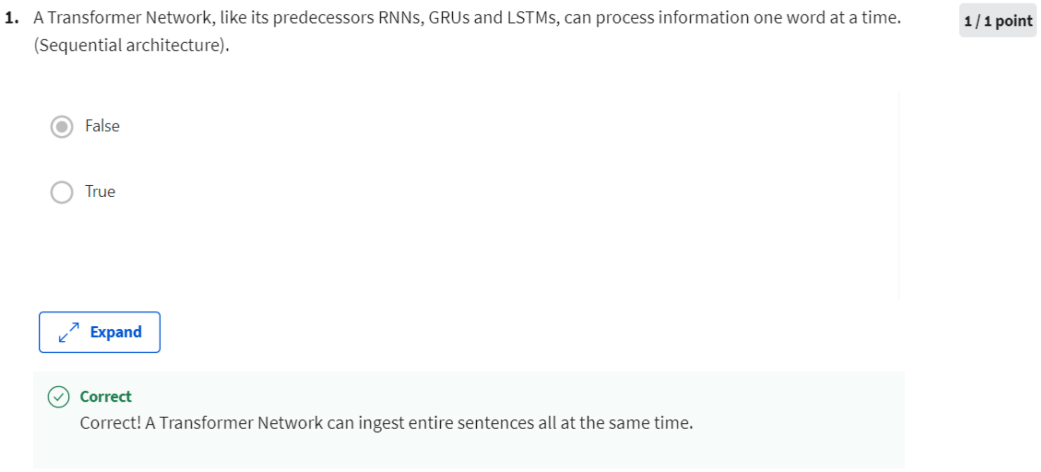
Transformers



1. A Transformer Network, like its predecessors RNNs, GRUs and LSTMs, can process information one word at a time. (Sequential architecture).

1 / 1 point

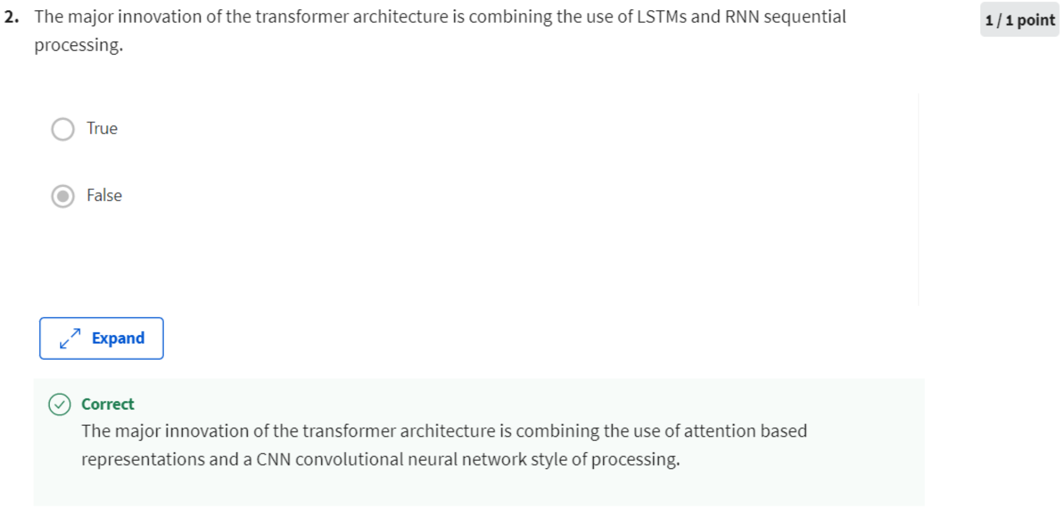
● False

○ True

Expand

✓ Correct

Correct! A Transformer Network can ingest entire sentences all at the same time.



2. The major innovation of the transformer architecture is combining the use of LSTMs and RNN sequential processing.

1 / 1 point

○ True

● False

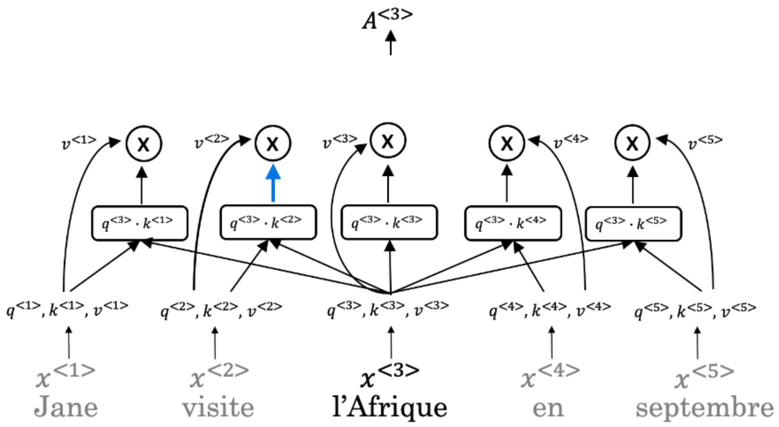
Expand

✓ Correct

The major innovation of the transformer architecture is combining the use of attention based representations and a CNN convolutional neural network style of processing.

3. The concept of *Self-Attention* is that:

1 / 1 point



○ Given a word, its neighbouring words are used to compute its context by selecting the highest of those word values to map the Attention related to that given word.

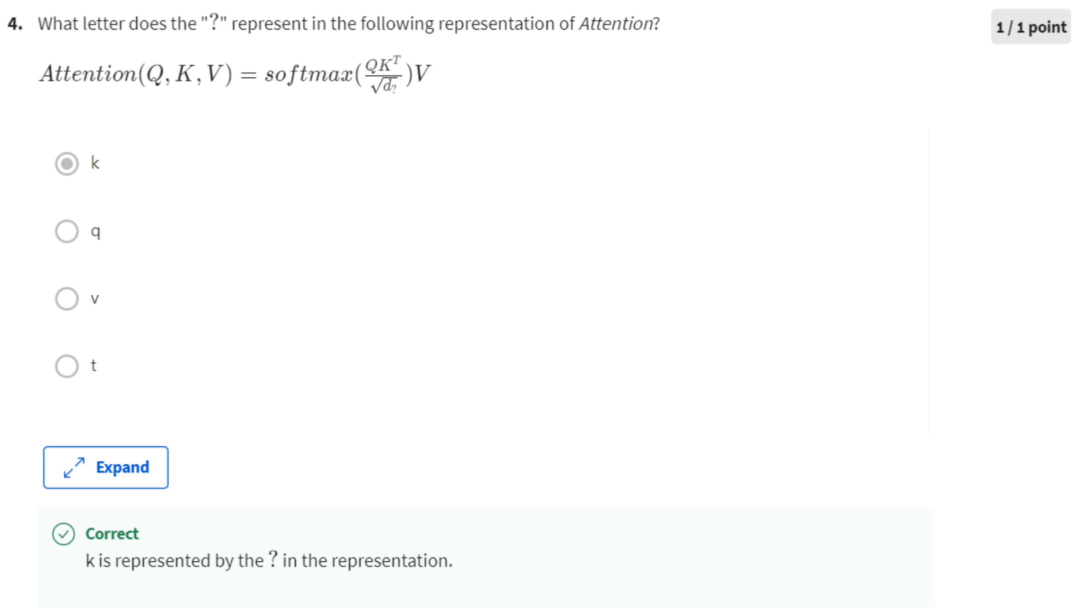
○ Given a word, its neighbouring words are used to compute its context by taking the average of those word values to map the Attention related to that given word.

● Given a word, its neighbouring words are used to compute its context by summing up the word values to map the Attention related to that given word.

○ Given a word, its neighbouring words are used to compute its context by selecting the lowest of those word values to map the Attention related to that given word.

Expand

✓ Correct



4. What letter does the "$?$" represent in the following representation of *Attention*?

1 / 1 point

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_?}})V$$

● k

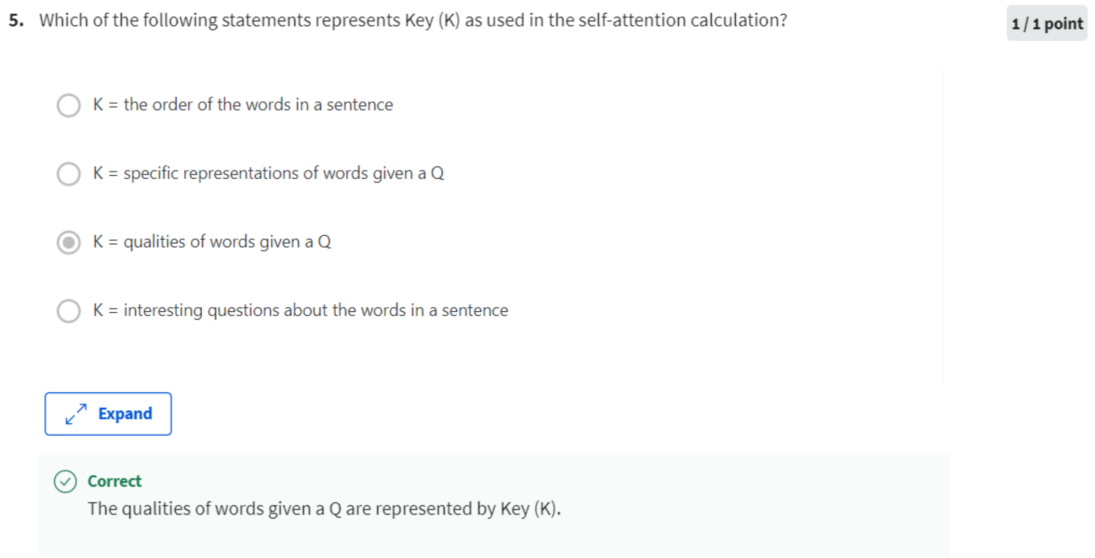
○ q

○ v

○ t

Expand

✓ **Correct**
k is represented by the $?$ in the representation.



5. Which of the following statements represents Key (K) as used in the self-attention calculation?

1 / 1 point

○ K = the order of the words in a sentence

○ K = specific representations of words given a Q

● K = qualities of words given a Q

○ K = interesting questions about the words in a sentence

Expand

✓ **Correct**
The qualities of words given a Q are represented by Key (K).

6. *Attention*$(W_i^Q Q, W_i^K K, W_i^V V)$

1 / 1 point

What does i represent in this multi-head attention computation?

○ The computed attention weight matrix associated with specific representations of words given a Q

● The computed attention weight matrix associated with the *ith* “head” (sequence)

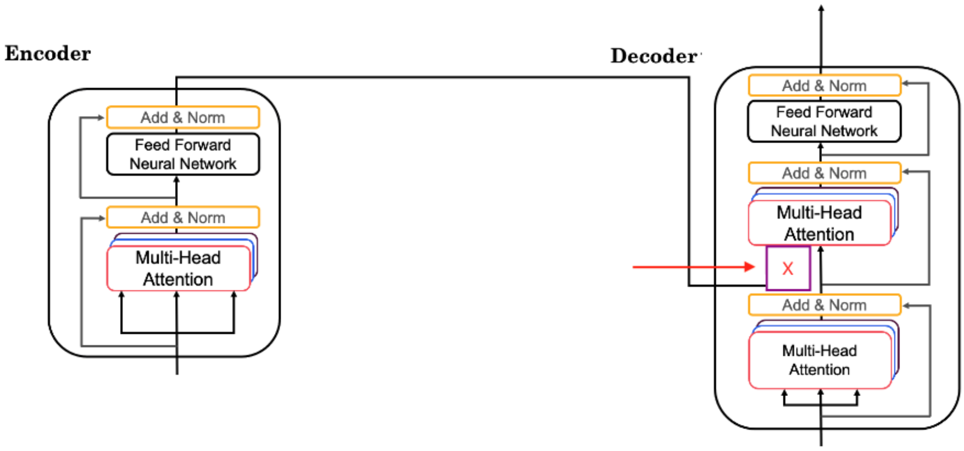
○ The computed attention weight matrix associated with the *ith* “word” in a sentence.

○ The computed attention weight matrix associated with the order of the words in a sentence

Expand

Correct
i here represents the computed attention weight matrix associated with the “head” (sequence).

7. Following is the architecture within a Transformer Network ***(without displaying positional encoding and output layers(s)).***



What information does the *Decoder*take from the *Encoder* for its second block of *Multi-Head Attention* ? (Marked X, pointed by the independent arrow)

What information does the *Decoder*take from the *Encoder* for its second block of *Multi-Head Attention*? (Marked X, pointed by the independent arrow)

(Check all that apply)

□ Q

☑ V

✓ Correct

☑ K

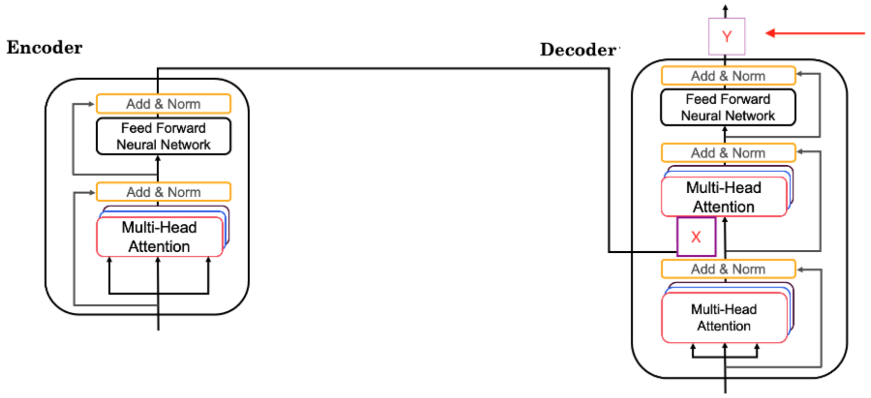
✓ Correct

Expand

✓ **Correct**
Great, you got all the right answers.

8. Following is the architecture within a Transformer Network. ***(without displaying positional encoding and output layers(s))***

1 / 1 point



What is the output layer(s) of the *Decoder*? (Marked Y, pointed by the independent arrow)

What is the output layer(s) of the *Decoder*? (Marked Y, pointed by the independent arrow)

● Linear layer followed by a softmax layer.

○ Linear layer

○ Softmax layer followed by a linear layer.

○ Softmax layer

Expand

✓ **Correct**

9. Which of the following statements is true about positional encoding? Select all that apply.

0 / 1 point

☐ Positional encoding is used in the transformer network and the attention model.

☑ Positional encoding is important because position and word order are essential in sentence construction of any language.

✓ **Correct**
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

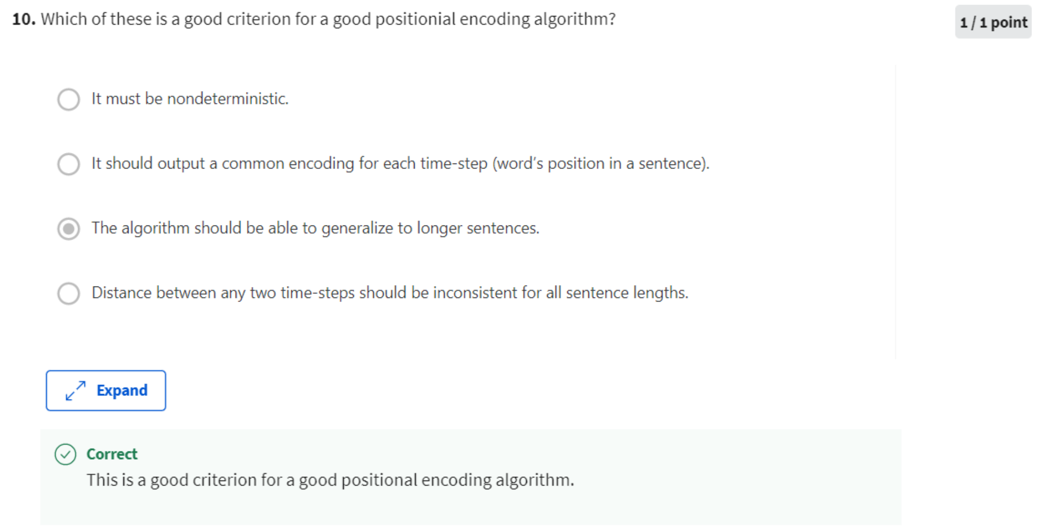
☐ Positional encoding uses a combination of sine and cosine equations.

☑ Positional encoding provides extra information to our model.

✓ **Correct**
This is a correct answer, but other options are also correct. To review the concept watch the lecture *Transformer Network*.

Expand

☒ **Incorrect**
You didn't select all the correct answers



10. Which of these is a good criterion for a good positionial encoding algorithm?

1 / 1 point

○ It must be nondeterministic.

○ It should output a common encoding for each time-step (word's position in a sentence).

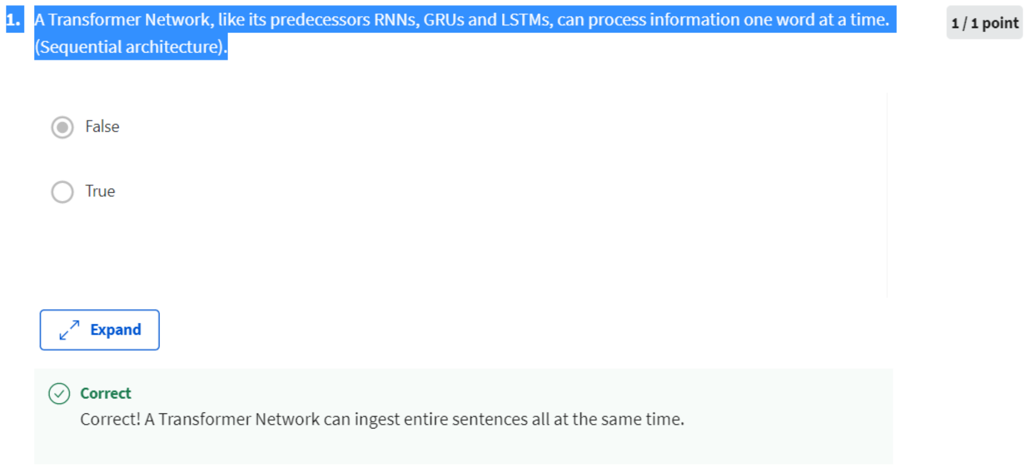
● The algorithm should be able to generalize to longer sentences.

○ Distance between any two time-steps should be inconsistent for all sentence lengths.

Expand

✓ **Correct**

This is a good criterion for a good positional encoding algorithm.



1. A Transformer Network, like its predecessors RNNs, GRUs and LSTMs, can process information one word at a time. (Sequential architecture).

1 / 1 point

● False

○ True

Expand

✓ **Correct**

Correct! A Transformer Network can ingest entire sentences all at the same time.

2. Transformer Network methodology is taken from:

1 / 1 point

● Attention Mechanism and CNN style of processing.

○ Attention Mechanism and RNN style of processing.

○ RNN and LSTMs

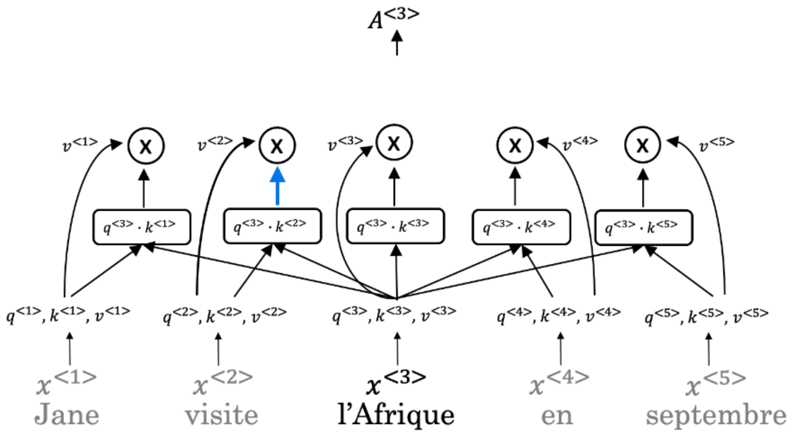
○ GRUs and LSTMs

Expand

Correct
Transformer architecture combines the use of attention based representations and a CNN convolutional neural network style of processing.

3. The concept of *Self-Attention* is that:

1 / 1 point



○ Given a word, its neighbouring words are used to compute its context by selecting the highest of those word values to map the Attention related to that given word.

● Given a word, its neighbouring words are used to compute its context by summing up the word values to map the Attention related to that given word.

○ Given a word, its neighbouring words are used to compute its context by taking the average of those word values to map the Attention related to that given word.

○ Given a word, its neighbouring words are used to compute its context by selecting the lowest of those word values to map the Attention related to that given word.

Expand

✓ Correct

4. Which of the following correctly represents *Attention ?*

1 / 1 point

● $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

○ $Attention(Q, K, V) = min(\frac{QK^T}{\sqrt{d_k}})V$

○ $Attention(Q, K, V) = min(\frac{QV^T}{\sqrt{d_k}})K$

Expand

✓ Correct

5. Are the following statements true regarding Query (Q), Key (K) and Value (V)?

1 / 1 point

Q = interesting questions about the words in a sentence

K = qualities of words given a Q

V = specific representations of words given a Q

○ False

● True

Expand

Correct
Q = interesting questions about the words in a sentence, K = qualities of words given a Q, V = specific representations of words given a Q

6. $\boldsymbol{Attention(W_i^Q Q, W_i^K K, W_i^V V)}$

1 / 1 point

What does i represent in this multi-head attention computation?

○ The computed attention weight matrix associated with the order of the words in a sentence

○ The computed attention weight matrix associated with specific representations of words given a Q

○ The computed attention weight matrix associated with the *ith* "word" in a sentence.

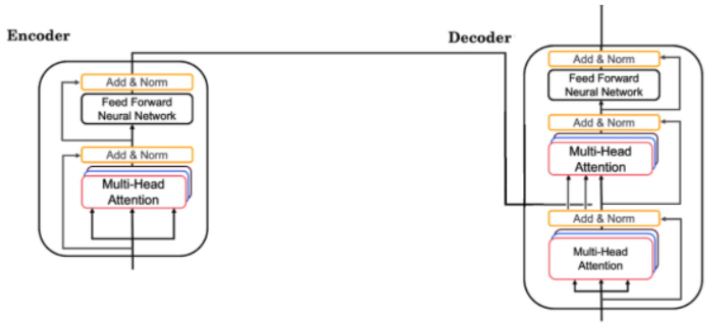
● The computed attention weight matrix associated with the *ith* "head" (sequence)

Expand

Correct
i here represents the computed attention weight matrix associated with the "head" (sequence).

7. Following is the architecture within a Transformer Network ***(without displaying positional encoding and output layers(s)).***

0 / 1 point



What is generated from the output of the *Decoder's* first block of *Multi-Head Attention*?

● V

○ Q

○ K

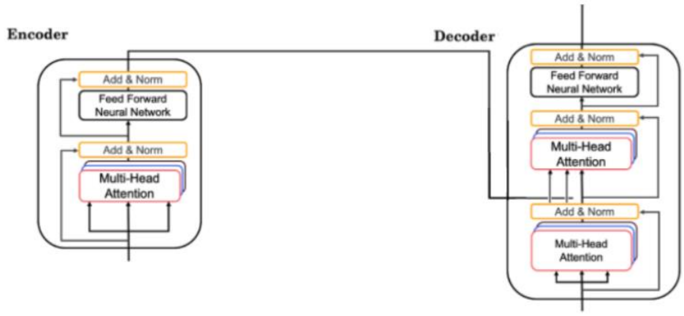
Expand

Incorrect

To revise the concept watch the lecture .

8. Following is the architecture within a Transformer Network ***(without displaying positional encoding and output layers(s)).***

1 / 1 point



The output of the decoder block contains a softmax layer followed by a linear layer to predict the next word one word at a time.

The output of the decoder block contains a softmax layer followed by a linear layer to predict the next word one word at a time.

○ True

● False

Expand

Correct

The output of the decoder block contains a linear layer followed by a softmax layer to predict the next word one word at a time.

9. Which of the following statements is true?

0 / 1 point

○ The transformer network differs from the attention model in that only the transformer network contains positional encoding.

○ The transformer network is similar to the attention model in that neither contain positional encoding.

○ The transformer network differs from the attention model in that only the attention model contains positional encoding.

● The transformer network is similar to the attention model in that both contain positional encoding.

Expand

⊗ **Incorrect**

To revise the concept watch the lecture .

10. Which of these is ***not*** a good criterion for a good positional encoding algorithm?

0 / 1 point

○ It should output a common encoding for each time-step (word's position in a sentence).

○ Distance between any two time-steps should be consistent for all sentence lengths.

○ The algorithm should be able to generalize to longer sentences.

● It must be deterministic.

Expand

⊗ **Incorrect**

This is a good criterion for a good positional encoding algorithm.