# Evaluation process: Round 1

## For 128 tokens:

* **Accuracy: 0.476 (47.6%) - Indicates that 47.6% of the evaluations were correctly predicted as 'Correct'.**
* Precision: 1.0 (100%) -Precision being 1.0 means that all instances predicted as 'Correct' were actually 'Correct'.
* **Recall: 0.476 (47.6%) - Recall being 0.476 means that out of all actual 'Correct' instances, 47.6% were correctly identified.**
* F1 Score: 0.645 - This is the harmonic mean of precision and recall, providing a balance between the two.
* Confusion Matrix: Indicates that there are no true negatives or false positives, with 40 true positives and 44 false negatives.

## For 256 tokens:

* **Accuracy: 0.679 (67.9%) - Indicates that 67.9% of the evaluations were correctly predicted as 'Correct'.**
* Precision: 1.0 (100%) -Precision being 1.0 means that all instances predicted as 'Correct' were actually 'Correct'.
* **Recall: 0.679 (67.9%) - Recall being 0.679 means that out of all actual 'Correct' instances, 67.9% were correctly identified.**
* F1 Score: 0.809 - This is the harmonic mean of precision and recall, providing a balance between the two.
* Confusion Matrix: Indicates that there are no true negatives or false positives, with 57 true positives and 27 false negatives.

In our case, only accuracy and recall provide meaningful insights. In our testing scenario, accuracy and recall will always be identical because all answers must be correct, with no incorrect responses allowed.

What we can see here is that with 128 tokens **47.6%** of the questions in the testing set are correctly answered by the model. That is not good enough as over the half of the questions are not correctly answered.

With 256 tokens we can see a lot of improvement as the model answered correctly **67.9%** of the questions.

**(Note: NotFull answers considered Incorrect for the evaluation process)**

## Ideas for improvement:

1. Try run the model with **512 tokens** as it seems that more tokens help the model to form and return correct answers
2. See the **articles again** and verify that all the necessary information is correctly formed and clearly presented in each article.

# Evaluation process: Round 2 – 512 tokens

## For 512 tokens:

* **Accuracy: 0.8452 (84.52%) - Indicates that 84.52%of the evaluations were correctly predicted as 'Correct'.**
* Precision: 1.0 (100%) -Precision being 1.0 means that all instances predicted as 'Correct' were actually 'Correct'.
* **Recall: 0.8452 (84.52%) - Recall being 0.8452 means that out of all actual 'Correct' instances, 84.52% were correctly identified.**
* F1 Score: 0.845- This is the harmonic mean of precision and recall, providing a balance between the two.
* Confusion Matrix: Indicates that there are no true negatives or false positives, with 71 true positives and 13 false negatives.

Comments: As we can see the first suggested improvement actually gives as much better results with the accuracy being 84.52% instead of 67.9%.

## Ideas for improvement:

1. See the **articles again** and verify that all the necessary information is correctly formed and clearly presented in each article.
2. Try **different prompts**. By refining the prompt it may help the model to provide more accurate and useful answers to the questions