Precision

- Of the chunks we retrieved, how many were correct?
- Interpretation:
  - High precision indicates an efficient system with few false positives.
  - Low precision suggests many irrelevant chunks are being retrieved.

$$\text{Precision} = \frac{\text{True Positives}}{\text{Total Retrieved}} = \frac{|\text{Retrieved} \cap \text{Correct}|}{|\text{Retrieved}|}$$

Recall

- Of all the correct chunks that exist, how many did we manage to retrieve?
- Interpretation:
  - High recall indicates comprehensive coverage of necessary information.
  - Low recall suggests important chunks are being missed.

$$\text{Recall} = \frac{\text{True Positives}}{\text{Total Correct}} = \frac{|\text{Retrieved} \cap \text{Correct}|}{|\text{Correct}|}$$

F1 Score

- The F1 score provides a balanced measure between precision and recall.
- It's the harmonic mean of precision and recall, tending towards the lower of the two values.
- Interpretation:
  - F1 score ranges from 0 to 1
  - An F1 score of 1.0 indicates perfect precision and recall.
  - An F1 score of 0.0 indicates the worst performance.
  - Generally, the higher the F1 score, the better the overall performance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

MRR(Mean Reciprocal Rank)

- MRR measures how well our system ranks relevant information.
- Interpretation:
  - MRR ranges from 0 to 1, where 1 is perfect (correct answer always first).
  - It only considers the rank of the first correct result for each query.
  - Higher MRR indicates better ranking of relevant information.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$