# BERTScore

## 🔹 BERTScore: Definition

BERTScore is a modern evaluation metric for text generation that uses pre-trained contextual embeddings
from models like BERT to compute semantic similarity between reference and candidate sentences.
Unlike traditional n-gram methods, BERTScore compares the meaning of words using cosine similarity between token embeddings.

## 🔸 BERTScore Formula

Let:
- R = Reference sentence
- C = Candidate sentence
- $e(r_i)$, $e(c_j)$ = contextual embeddings of token $r_i$ from reference and $c_j$ from candidate

Steps:
1. Compute cosine similarity between all token pairs: $\cos(e(c_j), e(r_i))$

Then:
- Precision (P): average maximum similarity for each token in candidate
  $P = (1 / |C|) * \sum_{c_j} (\max_{r_i} \cos(e(c_j), e(r_i)))$

- Recall (R): average maximum similarity for each token in reference
  $R = (1 / |R|) * \sum_{r_i} (\max_{c_j} \cos(e(r_i), e(c_j)))$

- F1-score (BERTScore):
  $F1 = (2 * P * R) / (P + R)$

## 🔹 Example

Reference: "a cat is sitting on the mat"

Candidate: "cat is sitting on mat"

### Cosine Similarity Matrix:

| | a | cat | is | sitting | on | the | mat |
|---|---|---|---|---|---|---|---|
| cat | 0.3 | 0.95 | 0.4 | 0.2 | 0.1 | 0.3 | 0.5 |
| is | 0.2 | 0.5 | 0.9 | 0.4 | 0.1 | 0.2 | 0.3 |
| sitting | 0.1 | 0.3 | 0.4 | 0.95 | 0.6 | 0.2 | 0.3 |
| on | 0.1 | 0.2 | 0.3 | 0.6 | 0.9 | 0.4 | 0.2 |

mat 0.2 0.4 0.3 0.1 0.2 0.5 0.93

### 🔸 Step-by-Step Calculation

1. Precision (Candidate → Reference)

Max similarities: [0.95, 0.9, 0.95, 0.9, 0.93]

P = (0.95 + 0.9 + 0.95 + 0.9 + 0.93) / 5 = 0.926

2. Recall (Reference → Candidate)

Max similarities: [0.3, 0.95, 0.9, 0.95, 0.9, 0.5, 0.93]

R = (0.3 + 0.95 + 0.9 + 0.95 + 0.9 + 0.5 + 0.93) / 7 ≈ 0.776

3. F1 Score (BERTScore)

F1 = (2 * 0.926 * 0.776) / (0.926 + 0.776) ≈ 0.844

### Final BERTScore ≈ 0.844

### 🔸 Summary Table

| Metric | Value |
|---|---|
| Precision (P) | 0.926 |
| Recall (R) | 0.776 |
| BERTScore (F1) | 0.844 |