**ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**

**Example**

**🔹 Reference Summary (Target):**

"The cat sat on the mat."

**🔹 Generated Summary (Prediction):**

"The cat lay on the rug."

**ROUGE-1**

**Unigrams in the Reference:**

["the", "cat", "sat", "on", "the", "mat"]

(Note: duplicate words are counted)

**Unigrams in the Generated:**

["the", "cat", "lay", "on", "the", "rug"]

**Overlap Unigrams:**

* "the" (counts: 2 in both)
* "cat"
* "on"

✅ Overlap: **"the" (2 times), "cat", "on"** → Total = **4 overlapping unigrams**

**📈 ROUGE-1 Scores**

Let’s compute **Precision**, **Recall**, and **F1-score**:

* **Precision** = (Overlapping unigrams) / (Total unigrams in generated summary)  
  = 4 / 6 ≈ **0.667**
* **Recall** = (Overlapping unigrams) / (Total unigrams in reference summary)  
  = 4 / 6 ≈ **0.667**
* **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall)  
  = 2 × (0.667 × 0.667) / (0.667 + 0.667) ≈ **0.667**

**ROUGE-2**

checks for overlapping **bigrams** (two-word sequences).

**✅ Reference Bigrams:**

* "the cat"
* "cat sat"
* "sat on"
* "on the"
* "the mat"

**✅ Generated Bigrams:**

* "the cat"
* "cat lay"
* "lay on"
* "on the"
* "the rug"

**📊 Overlapping Bigrams:**

* "the cat" ✅
* "on the" ✅

**Overlap Count**: 2

**📈 ROUGE-2 Scores:**

* **Precision** = 2 / 5 = **0.4**
* **Recall** = 2 / 5 = **0.4**
* **F1-score** = 2 × (0.4 × 0.4) / (0.4 + 0.4) = **0.4**

**ROUGE-L**

measures the **longest subsequence of words** that appear in both the reference and generated summary in the same order (but not necessarily contiguous).

**Reference: ["the", "cat", "sat", "on", "the", "mat"]**

**Generated: ["the", "cat", "lay", "on", "the", "rug"]**

**Longest Common Subsequence (LCS):**

* **"the", "cat", "on", "the"** → Length = **4**

**ROUGE-L Scores:**

* **Precision** = LCS length / Generated length = 4 / 6 ≈ **0.667**
* **Recall** = LCS length / Reference length = 4 / 6 ≈ **0.667**
* **F1-score** = 2 × (0.667 × 0.667) / (0.667 + 0.667) ≈ **0.667**