ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation – N-gram variant) is a widely used metric in **Natural Language Processing (NLP)** for evaluating **automatically generated text**, such as summaries, translations, or paraphrases. It works by measuring **n-gram overlap** between the candidate (generated) text and one or more reference (human-written) texts.

An **n-gram** is a contiguous sequence of *n* words from a given text:

- **Unigram** (n=1): individual words
- **Bigram** (n=2): two-word sequences
- **Trigram** (n=3), etc.

ROUGE-N specifically compares the n-grams of the candidate and reference texts to evaluate how much they overlap. The metric is **recall-oriented**, but can also compute **precision** and **F1-score**.

ROUGE-N Recall:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram} \in \text{ref}} \text{Count}_{\text{match}}(\text{gram})}{\sum_{\text{gram} \in \text{ref}} \text{Count}(\text{gram})}$$

Precision:

$$\frac{\text{Number of overlapping n-grams}}{\text{Total n-grams in candidate}}$$

F1 Score:

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example

- **Reference**: "The cat sits on the mat"
- **Candidate**: "The cat sits on the floor"
- **ROUGE-1 (unigram)** overlap: “The”, “cat”, “sits”, “on” → 4 matches
- **ROUGE-2 (bigram)** overlap: “The cat”, “cat sits”, “sits on”, “on the” → matches “The cat”, “cat sits”, “sits on” → 3 matches

Interpretation

- **High ROUGE-N** → High similarity to reference
- **Low ROUGE-N** → Low content overlap
- **Higher N (ROUGE-2, ROUGE-3...)** → More syntactic/phrase-level fidelity

Variants

- **ROUGE-1**: Focuses on individual word overlap
- **ROUGE-2**: Captures short phrase overlap
- **ROUGE-L**: Longest Common Subsequence
- **ROUGE-S**: Skip-bigram-based variant

Applications

- **Text Summarization**: Evaluating how well a generated summary captures reference content.
- **Machine Translation**: Assessing translation quality via n-gram overlap with a human reference.
- **Paraphrase Generation**: Measuring the similarity of alternative wordings.
- **Dialogue Systems**: Evaluating chatbot or dialogue responses.

Limitations

- ROUGE only measures **surface-level overlap** and does not consider **semantic similarity**.
- It may penalize **paraphrased** but semantically correct outputs.
- Best used with **multiple reference texts** for fair evaluation.