1. Evaluation Metrics

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

ROUGE is a recall oriented evaluation metric used in summarization and text generation tasks. It measures how much of the reference text's content is captured by the system-generated text based on word overlap.

Types of ROUGE:

- **ROUGE-N:** Measures the overlap of n-grams between the candidate and the reference.
  - ROUGE-1 (unigrams) captures basic lexical overlap.
  - ROUGE-2 (bigrams) captures short phrase overlap and fluency.
- **ROUGE-L:** Based on the Longest Common Subsequence (LCS) between candidate and reference. It accounts for sentence-level structure and fluency, not just word overlap.

Importance:
ROUGE is easy to compute, language independent, and interpretable. It aligns well with human judgments of content coverage in summarization. Since it emphasizes recall, it rewards systems that include more of the reference's key information.

Drawbacks:
ROUGE only considers exact word matches, ignoring synonyms and paraphrases. It fails to measure factual accuracy or coherence and may reward verbose outputs that copy large parts of the reference. Therefore, while ROUGE is useful for recall-based evaluation, it does not assess meaning preservation or readability effectively.

(b) BLEU (Bilingual Evaluation Understudy)

BLEU is a precision based metric originally developed for machine translation. It evaluates how many n-grams in the system output appear in one or more reference translations. BLEU computes modified n-gram precision for n = 1 to 4 and combines them using a geometric mean. A brevity penalty is applied to penalize excessively short translations.

Importance:
It provides stable results at the corpus level and enables quick comparison between systems. Its reliance on multiple n-gram precisions helps ensure both accuracy and fluency to some extent.

Drawbacks:
BLEU focuses only on surface level overlap, ignoring meaning similar to ROGUE. Two semantically equivalent sentences using different words can receive a low BLEU score. It also correlates poorly with human judgments at the sentence level and struggles with free-form generation tasks like dialogue or summarization.

(c) BERTScore

BERTScore is a semantic similarity based metric that uses contextual embeddings from models like BERT or RoBERTa. Instead of comparing words directly, it compares vector representations of tokens in the candidate and reference sentences.

How it works:
Both texts are encoded using a pretrained language model. Pairwise cosine similarity is computed between all token embeddings. For each token in one text, the most similar token in the other text is found. These similarities are averaged to produce precision, recall, and F1 scores.

Importance:
BERTScore captures meaning rather than exact word overlap. It recognizes synonyms and paraphrases, making it more aligned with human judgment. It is particularly effective for evaluating summarization, dialogue, and open ended generation tasks.

Drawbacks:
It is computationally heavy and depends on the quality and domain of the underlying language model. Results can vary across models and languages. Additionally, BERTScore's numeric values (usually around 0.8–0.9) are less intuitive to interpret compared to BLEU or ROUGE percentages.

---

2. Reference-Free Evaluations

Reference-free metrics evaluate generated text without requiring human written references. They assess intrinsic qualities such as fluency, diversity, or coherence. These methods are valuable in open-ended tasks (e.g., dialogue generation, creative writing) where multiple valid outputs exist.

Example – Self-BLEU

Definition:
Self-BLEU is a modification of BLEU used to measure diversity among multiple generated outputs. Each generated sentence is treated as a candidate, and all other sentences are treated as references. BLEU is then calculated for each, and the results are averaged.

Interpretation:

- **High Self-BLEU:** Outputs are very similar implies low diversity (model repeats itself).
- **Low Self-BLEU:** Outputs differ more implies high diversity (model is creative).

Advantages:

- Does not require human references.
- Useful for detecting mode collapse in generative models (when the model produces repetitive outputs or hallucinations).
- Works across domains and languages.

Disadvantages:

- Like BLEU, it focuses on lexical overlap, not meaning.
- High diversity (low Self-BLEU) doesn't always mean high quality or factuality.
- Does not measure relevance or correctness relative to the input or task.