Understanding LLM Evaluation Metrics

1. Perplexity

- **Measures:** How well a language model predicts words.
- **Intuition:** Lower perplexity → model predicts text better.
- **Formula (simplified):**

$$PP = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_1,...,w_{i-1})}$$

- **Use case:** Language modeling, text generation.
- **Interpretation:** Lower is better.

---

2. BLEU

- **Measures:** Precision of n-grams between model output and reference.
- **Focus:** How much of what the model wrote is correct.
- **Range:** 0 → 1 (or 0–100).
- **Use case:** Translation, summarization.
- **Limitation:** Doesn't care if model misses words.

Example: - Reference: "The cat sat on the mat" - Model: "The cat is on the mat" - BLEU checks **matching n-grams**.

---

3. ROUGE

- **Measures:** Recall of n-grams / sequences from reference.
- **Focus:** How much of the reference the model covered.
- **Variants:** ROUGE-N, ROUGE-L, ROUGE-S
- **Range:** 0 → 1 (or 0–100).
- **Use case:** Summarization, text coverage.

Example: - Reference: "The cat sat on the mat" - Model: "Cat on mat" - ROUGE checks **coverage of reference words**.

---

4. Accuracy / Task-Specific Metrics

- **Measures:** Correctness of output.
- **Use case:** QA, classification, code generation.
- **Interpretation:** Higher is better.

BLEU vs ROUGE – Simple Difference

Metric	Focus	Question it asks	Example
BLEU	Precision	“Of all words I wrote, how many are correct?”	Counts correct words in model output
ROUGE	Recall	“Of all words in reference, how many did I cover?”	Checks if all important reference words are included

Tip: - **BLEU = Don’t write wrong words** - **ROUGE = Don’t miss important words**