Q9. List the Different Evaluation Metrics for Language Models

Evaluating a language model helps determine how well it can predict, generate, or classify text. The choice of evaluation metric depends on the **task type** (e.g., language modeling, translation, classification).

🔹 1. Perplexity

✅ Definition:

- Perplexity measures how well a language model predicts a sequence of words.
- Lower perplexity = better prediction performance.

📌 Formula:

$$\text{Perplexity} = 2^{H(P)} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)}$$

- $P(w_i)$: probability assigned to the i-th word by the model
- N: total number of words

💡 Intuition:

- Perplexity is the exponentiated average negative log-likelihood.
- A perplexity of 100 means the model is as uncertain as randomly choosing among 100 options at each step.

🧠 Example:

If a sentence has 10 words and the model assigns good probabilities to each, it might have perplexity ~30. If it performs poorly, perplexity could rise to ~100+.

🔹 2. Cross-Entropy

✅ Definition:

- Cross-entropy measures the difference between the predicted probability distribution and the actual (true) distribution.

📌 Formula:

$$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

- $P(x)$: true distribution
- $Q(x)$: predicted distribution

💡 Intuition:

- Lower cross-entropy indicates that the model's predictions are closer to the actual outcomes.
- It is directly related to perplexity (Perplexity = $2^{\text{Cross-Entropy}}$).

🔹 3. Accuracy

✅ **Definition:**

- Used in classification tasks such as POS tagging or sentence classification.
- Measures the proportion of correctly predicted outputs.

📌 **Formula:**

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

🧠 **Example:**

In sentiment classification, if the model correctly labels 950 out of 1000 sentences, accuracy = 95%.

🔹 4. BLEU Score (Bilingual Evaluation Understudy)

✅ **Definition:**

- A metric for evaluating **text generation** tasks such as **machine translation**, **summarization**, or **captioning**.
- Compares overlap between generated and reference n-grams.

📌 **Formula (Simplified):**

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- p_n: precision for n-grams (1-gram, 2-gram...)
- BP: brevity penalty (prevents overly short translations)
- w_n: weight for each n-gram level (usually equal)

🧠 **Example:**

If a translation output shares many bigrams and trigrams with the reference sentence, BLEU will be high (close to 1.0).

🔹 5. Precision, Recall, F1 Score

✅ **Usage:**

- Especially important in **sequence labeling**, **NER**, or **text classification**.

📌 **Formulas:**

- **Precision** = $\frac{TP}{TP+FP}$
- **Recall** = $\frac{TP}{TP+FN}$
- **F1 Score** = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$

🧠 **Example:**

In named entity recognition (NER), if the model identifies “New York” as a location correctly, it's a True Positive.

🔹 6. Log-Likelihood

✅ **Definition:**

- Sum of the log probabilities assigned to the correct words in a sequence.
- Often used during training as a loss function.

$$\log P(w_1, w_2, ..., w_N) = \sum_{i=1}^{N} \log P(w_i | w_{<i})$$

- Higher log-likelihood indicates better model fit.

🔹 7. Token-level Error Rate

- Counts how many individual tokens were predicted incorrectly.
- Useful for evaluating low-level models (e.g., in speech recognition or OCR).

🔹 8. Edit Distance / Levenshtein Distance

- Used in spelling correction or speech recognition.
- Measures how many operations (insertions, deletions, substitutions) are required to convert the output to the correct sequence.

✅ Summary: When to Use Which Metric

Metric	Use Case	Good Value
Perplexity	Language modeling	Lower is better
Cross-Entropy	Language modeling	Lower is better
Accuracy	Classification, tagging	Higher is better
BLEU Score	Translation, summarization	Closer to 1.0
Precision/Recall/F1	NER, sentiment, sequence labeling	Higher is better
Edit Distance	Spelling correction, transcription	Lower is better