A Comprehensive Reference on Evaluation Metrics for Large Language Models (LLMs)

Eng. Ahmed Métwalli

December 11, 2024

Contents

1 Introduction

This document provides a structured overview of commonly used evaluation metrics for Large Language Models (LLMs). Each metric is presented with:

- A brief **definition** and **use case**.
- A **mathematical formula**, where applicable.
- Detailed descriptions of each symbol in the formula.
- Considerations and limitations of using the metric.

By compiling these references in a single document, you can streamline the evaluation process and ensure clarity and rigor in assessing model performance.

2 Intrinsic Metrics

Intrinsic metrics evaluate the model's generated text without requiring a downstream task. They often compare the model's output against a reference dataset or assess internal predictive quality.

2.1 Perplexity (PPL)

Definition: Perplexity measures how well a language model predicts a sequence of tokens. Lower perplexity indicates better predictive performance and language fluency.

Use Case: Commonly used to evaluate language modeling tasks.

Formula:

$$\text{PPL} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_{<i})} \tag{1}$$

Symbols:

- N: Total number of tokens in the test set.
- w_i: The i-th token in the test sequence.
- $w_{<i}$: The sequence of tokens preceding w_i.
- $P(w_i \mid w_{<i})$: The conditional probability assigned by the model to the token w_i given the preceding context $w_{<i}$.

Interpretation: A lower PPL indicates the model's predictions align more closely with the test data distribution.

2.2 BLEU (Bilingual Evaluation Understudy)

Definition: BLEU compares the machine-generated text to one or more reference texts by measuring the overlap of n-grams.

Use Case: Machine translation, where one compares the generated translation to a reference human translation.

Formula (simplified):

$$\text{BLEU} = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \times \exp\left(\min\left(0, 1 - \frac{r}{c}\right)\right) \tag{2}$$

Symbols:

- N: Maximum n-gram length (commonly $N = 4$).
- w_n: Weight for the n-gram precision (often $w_n = \frac{1}{N}$ for uniform weighting).
- p_n: Modified precision for n-gram matches, i.e.,

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}(n\text{-gram})}$$

- r: Effective reference length (closest reference length to the candidate text).
- c: Length of the candidate text.

Interpretation: A higher BLEU score indicates a closer match to the reference. However, BLEU is sensitive to exact wording and does not measure semantic similarity well.

2.3 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Definition: ROUGE measures the overlap between a generated summary and a reference summary, focusing on recall of n-grams and sequences.

Use Case: Text summarization tasks.

Common Variants:

- ROUGE-1: Overlap of unigrams.
- ROUGE-2: Overlap of bigrams.
- ROUGE-L: Overlap based on longest common subsequence (LCS).

Example (ROUGE-1):

$$\text{ROUGE-1} = \frac{\sum_{\text{reference n-grams}} \min(\text{Count}_{\text{candidate}}(\text{1-gram}), \text{Count}_{\text{reference}}(\text{1-gram}))}{\sum_{\text{reference n-grams}} \text{Count}_{\text{reference}}(\text{1-gram})} \tag{3}$$

Symbols:

- $\text{Count}_{\text{candidate}}(\text{1-gram})$: Frequency of a specific unigram in the candidate summary.
- $\text{Count}_{\text{reference}}(\text{1-gram})$: Frequency of the same unigram in the reference summary.

Interpretation: Higher ROUGE scores suggest the candidate summary covers more of the same content as the reference. ROUGE is surface-level and does not guarantee semantic equivalence.

2.4 METEOR

Definition: METEOR uses a combination of exact word matches, stem matches, synonym matches, and paraphrase matches between candidate and reference texts.

Use Case: Machine translation, summarization. More semantically focused than BLEU.

Formula (simplified):

$$\text{METEOR} = 10 \times \frac{P \times R}{(R + P)} \times (1 - 0.5 \times (\text{frag}/\text{matchCount})) \tag{4}$$

Symbols:

- P: Precision of matched unigrams.
- R: Recall of matched unigrams.
- frag: Number of matched chunks (sequences of matched words).
- matchCount: Total number of matched unigrams.

Interpretation: METEOR accounts for more flexible matches than BLEU or ROUGE, aiming for a better correlation with human judgments.

2.5 BERTScore

Definition: BERTScore uses contextual embeddings (e.g., from BERT) to measure semantic similarity between candidate and reference sentences.

Use Case: Evaluation tasks where semantic equivalence is more important than exact word match.

Formula (high-level):

$$\text{BERTScore}(C, R) = \frac{1}{|C|} \sum_{c \in C} \max_{r \in R} \text{cosine}(E(c), E(r))$$

(And similarly for recall, then F1.)

Symbols:

- C: Set of tokens in the candidate text.
- R: Set of tokens in the reference text.
- $E(x)$: Embedding of token x using a model like BERT.
- $\text{cosine}(E(c), E(r))$: Cosine similarity between the embeddings of tokens c and r.

Interpretation: BERTScore aligns tokens semantically and allows different phrasing. A higher BERTScore indicates closer semantic equivalence.

3 Factual Consistency and Accuracy Metrics

3.1 Factuality Checks

Factuality checks compare model-generated statements against a trusted knowledge base (e.g., Wikipedia or Wikidata) or use specialized fact-checking models.

Approach:

- Compare asserted facts in generated text with known facts in a knowledge base.
- Use automatic fact-checking tools (like FEVER score) that evaluate claim veracity.

There is no single closed-form equation here; metrics vary by implementation.

4 Extrinsic Metrics

Extrinsic metrics evaluate how well the model performs on downstream tasks, often with a well-defined ground truth.

4.1 Exact Match (EM)

Definition: The percentage of generated answers that match the ground-truth answer exactly, character-for-character.

Use Case: Tasks like question answering (e.g., SQuAD).

Formula:

$$\text{EM} = \frac{\sum_{i=1}^{M} \mathbf{1}(\hat{y}_i = y_i)}{M} \tag{5}$$

Symbols:

- M: Number of test examples.
- $\hat{y}_i$: Predicted answer for the i-th example.
- y_i: Ground-truth answer for the i-th example.
- $\mathbf{1}(\cdot)$: Indicator function that is 1 if the condition is true, 0 otherwise.

4.2 F1-Score (For Classification or QA Overlap)

Definition: F1 is the harmonic mean of precision and recall, often used in classification or QA tasks where partial overlaps matter.

Formula:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Symbols:

- Precision $= \frac{TP}{TP+FP}$, where TP = True Positives, FP = False Positives.
- Recall $= \frac{TP}{TP+FN}$, where FN = False Negatives.

5 Efficiency Metrics

5.1 Latency

Definition: Time taken for the model to produce an output after receiving an input.

Interpretation: Lower latency is essential for real-time applications.

5.2 Throughput

Definition: Number of predictions the model can handle per unit time.

6 User-Centric Metrics

6.1 Human Evaluation

Definition: Human evaluators rate outputs for fluency, relevance, helpfulness, and factual correctness.

Approach:

- Use Likert scales (e.g., 1 to 5).
- Perform pairwise comparisons (A/B testing).

7 Conclusion

The choice of metric depends heavily on the task and the desired qualities in the LLM's output. For generative language models, a combination of intrinsic (e.g., Perplexity, BLEU, ROUGE, BERTScore) and extrinsic (e.g., EM, F1) metrics, as well as human evaluation, can provide a comprehensive picture of performance. Factuality checks and user-centric evaluations ensure that the model's outputs are trustworthy and align with user needs.

This reference document can guide you through selecting and interpreting appropriate metrics for any given scenario involving LLM evaluation.

Metric	Use Case	When Not to Use
Perplexity (PPL)	Evaluates language model fluency and internal consistency of predictions.	When the downstream task requires understanding beyond next-token prediction (e.g., factual correctness, reasoning).
BLEU	Machine translation quality, comparing output to a reference translation.	If exact lexical overlap is not crucial, or for tasks where semantic paraphrasing is acceptable and may be penalized.
ROUGE	Summarization tasks, measuring how much content in the reference is captured.	When you need to assess semantic similarity rather than lexical overlap (e.g., paraphrased summaries).
METEOR	Machine translation/summarization with some flexibility for synonyms and stems.	Tasks where deep semantic equivalence is crucial or where exact lexical matches are not relevant.
BERTScore	Evaluating semantic similarity between candidate and reference texts.	Pure lexical tasks or when you need a strict lexical match (BERTScore may be too lenient on rephrasings).
Exact Match (EM)	QA tasks with discrete answers where exact string match is required (e.g., closed-book QA).	Open-ended generative tasks, or when answers can be correct but phrased differently.
F1 Score	Classification or QA tasks where partial overlap matters (tokens or classes).	When lexical granularity is not important or when you need to capture deep semantic or factual correctness.
Factual Checks	Ensuring factual correctness against known knowledge bases.	Purely creative tasks where factual accuracy is not relevant, or tasks that do not rely on external knowledge.
Human Evaluation	Complex or open-ended tasks (dialogue, creativity, helpfulness).	Highly scalable automated testing scenarios, or when subjective human judgments are not viable.

Table 1: Overview of Evaluation Metrics, Their Use Cases, and When Not to Use Them