

Understanding Model Performance Metrics

Introduction

Assessing the performance of large language models involves evaluating their outputs against human-generated references. Traditional metrics like accuracy are insufficient due to the non-deterministic nature of language-based tasks. Instead, metrics like ROUGE and BLEU are employed to quantify the quality of generated text.

LLM Evaluation - Metrics



- Used for text summarization
- Compares a summary to one or more reference summaries



- Used for text translation
- Compares to human-generated translations

ROUGE Metric

ROUGE: **Recall-Oriented Understudy for Gisting Evaluation**

Usage: Evaluates the quality of automatically generated summaries by comparing them to human-generated references.

Terminology: Unigram (single word), Bigram (two words), N-gram (group of n words).

LLM Evaluation - Metrics - ROUGE-1

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1 Precision} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

$$\text{ROUGE-1 F1} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.8}{1.8} = 0.89$$

ROUGE-1: Measures unigram matches between the generated output and the reference.

LLM Evaluation - Metrics - ROUGE-2

Reference (human):
It is cold outside.

It is is cold
cold outside

Generated output:
It is very cold outside.

It is is very
very cold cold outside

$$\text{ROUGE-2 Recall:} = \frac{\text{bigram matches}}{\text{bigrams in reference}} = \frac{2}{3} = 0.67$$

$$\text{ROUGE-2 Precision:} = \frac{\text{bigram matches}}{\text{bigrams in output}} = \frac{2}{4} = 0.5$$

$$\text{ROUGE-2 F1:} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.335}{1.17} = 0.57$$

ROUGE-2: Considers bigram matches, accounting for word ordering.

LLM Evaluation - Metrics - ROUGE-L

Reference (human):
It is cold outside.

Generated output:
It is very cold outside.

$$\text{ROUGE-L Recall:} = \frac{\text{LCS}(\text{Gen}, \text{Ref})}{\text{unigrams in reference}} = \frac{2}{4} = 0.5$$

$$\text{ROUGE-L Precision:} = \frac{\text{LCS}(\text{Gen}, \text{Ref})}{\text{unigrams in output}} = \frac{2}{5} = 0.4$$

$$\text{ROUGE-L F1:} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.2}{0.9} = 0.44$$



ROUGE-L: Determines the longest common subsequence between the generated and reference outputs.

Limitations of ROUGE

Deceptive Scores: Simple ROUGE scores may not reflect subjective quality, allowing poor completions to yield high scores.

Clipping Function: Mitigates issues by limiting unigram matches to the maximum count in the reference.

LLM Evaluation - Metrics - ROUGE clipping

Reference (human): It is cold outside.	ROUGE-1 Precision = $\frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{4} = 1.0$	
Generated output: cold cold cold cold	Modified precision = $\frac{\text{clip(unigram matches)}}{\text{unigrams in output}} = \frac{1}{4} = 0.25$	
Generated output: outside cold it is	Modified precision = $\frac{\text{clip(unigram matches)}}{\text{unigrams in output}} = \frac{4}{4} = 1.0$	

BLEU Metric

BLEU: **B**ilingual **E**valuation **U**nderstudy

Usage: Evaluates the **quality of machine-translated text** by comparing it to human-generated translations.

Calculation: Averages precision across a range of n-gram sizes, comparing machine-generated and reference translations.

Example Evaluation

Reference Sentence: "I am very happy to say that I am drinking a warm cup of tea."

Candidate Sentences: Variations of the reference sentence with incremental modifications.

BLEU Scores: Range from 0 to 1, with higher scores indicating greater similarity to the reference.

Practical Considerations

Library Support: Pre-written libraries like Hugging Face facilitate easy calculation of ROUGE and BLEU scores.

Diagnostic Evaluation: ROUGE for summarization tasks, BLEU for translation tasks.

Comprehensive Evaluation: For a holistic assessment, leverage evaluation benchmarks developed by researchers.

Conclusion

ROUGE and BLEU metrics provide structured approaches to evaluate the performance of large language models. While simple and low-cost to calculate, they serve as diagnostic

tools rather than comprehensive evaluation measures. Researchers often rely on specialized evaluation benchmarks for a thorough assessment of model performance.