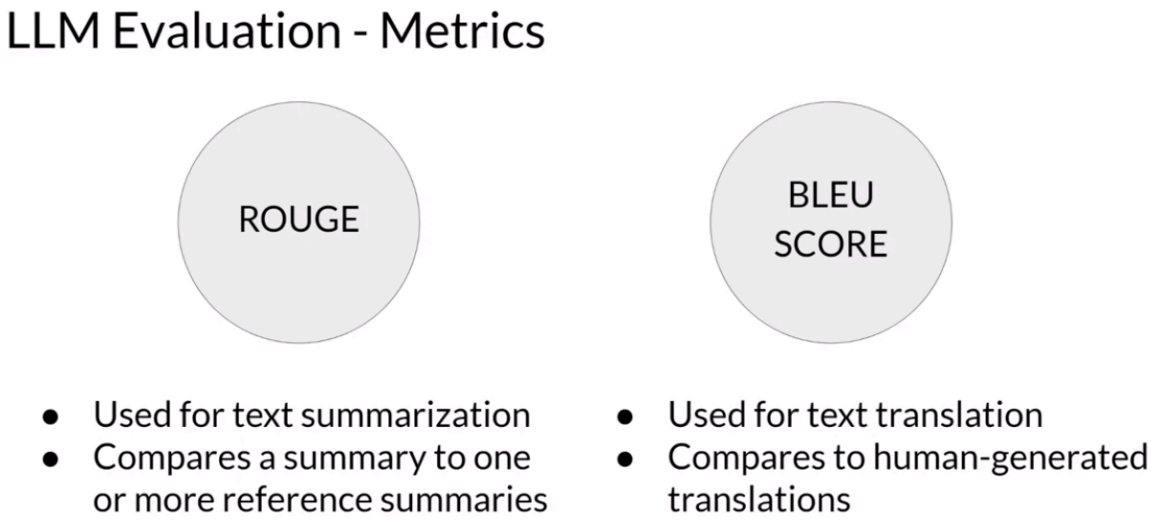
Understanding Model Performance Metrics

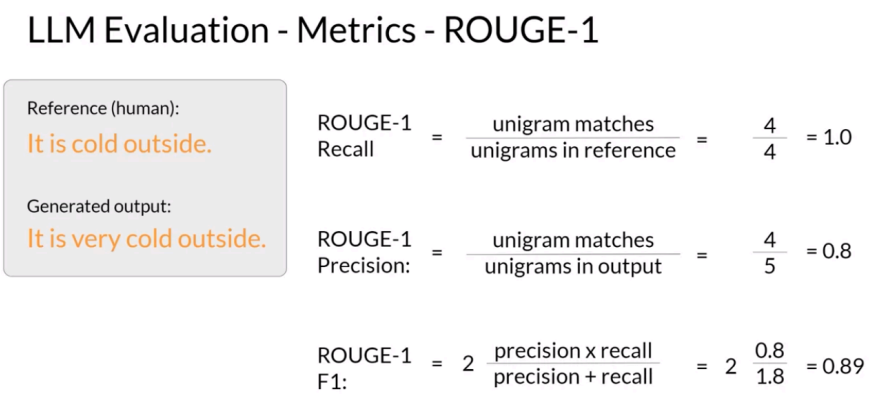
Introduction

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.

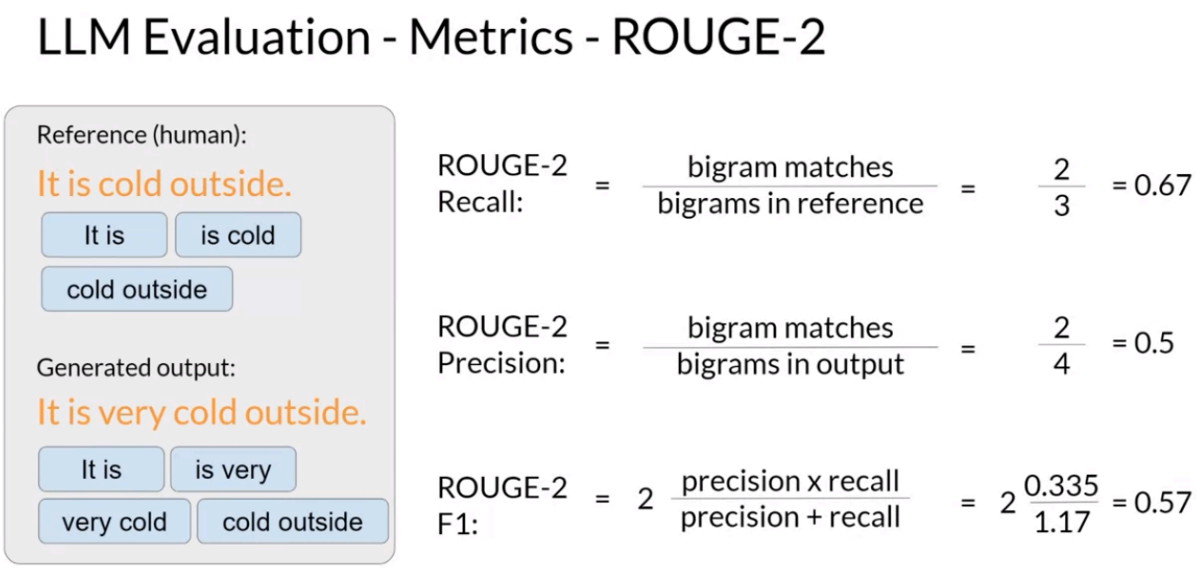


ROUGE Metric

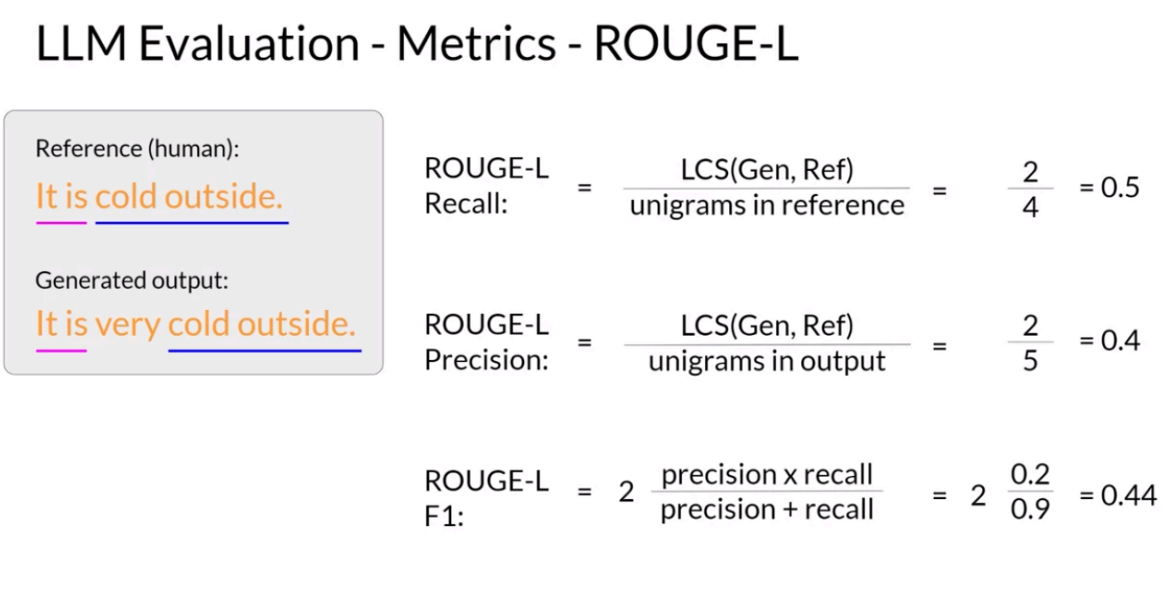
ROUGE: **Recall-Oriented Understudy for Gisting Evaluation**
Usage: Evaluates the quality of automatically generated summaries by comparing them to human-generated references.
Terminology: Unigram (single word), Bigram (two words), N-gram (group of n words).



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



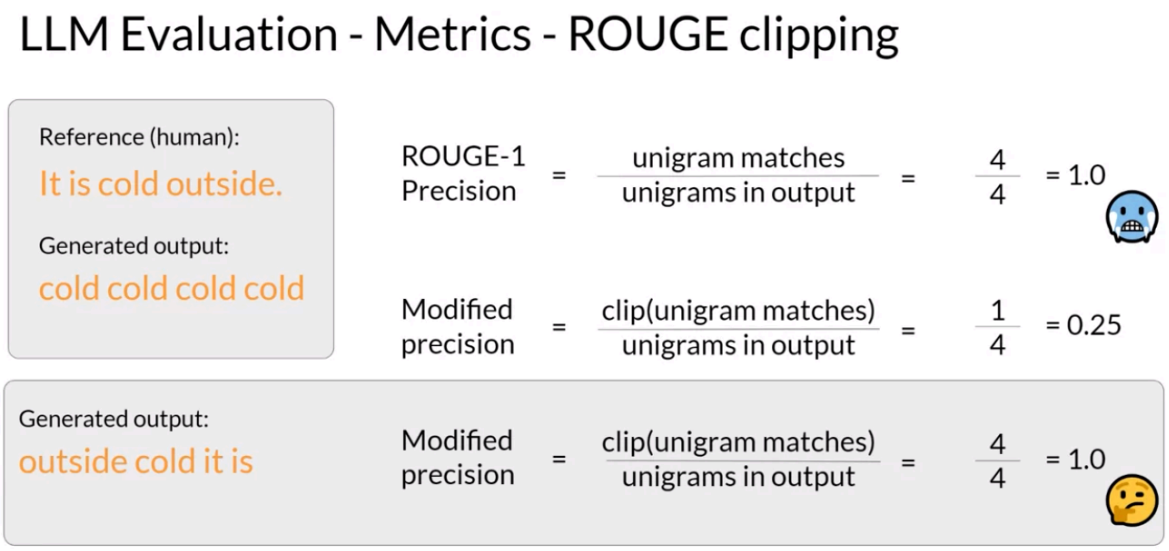
ROUGE-2: Considers bigram matches, accounting for word ordering.



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.



BLEU Metric

BLEU: **Bilingual Evaluation Understudy**
Usage: Evaluates the **quality of machine-translated tex**t 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.