

Metric

BLEU (bilingual evaluation understudy)

ROUGE (recall-oriented understudy for gisting evaluation)

BERTScore

LLM-as-judge

What it measures

N-gram overlap with reference text

Recall-oriented overlap with reference

Semantic similarity via embeddings

Quality scoring by a second LLM

Strengths

Fast, well-understood, language-agnostic

Good for summarization evaluation

Captures meaning beyond surface text

Evaluates nuance, reasoning, style

Limitations

Penalizes valid paraphrases; no semantic understanding

Same n-gram limitations as BLEU

Computationally expensive; requires embedding model

Expensive per evaluation; judge model has its own biases