Modern Evaluation Metrics Comparison (2024-2025)

Metric	What it Measures	Human Correlation	Cost	Best For
ROUGE	Word overlap	**0.45**	Free	Baseline screening
BLEU	N-gram precision	**0.38**	Free	Translation (not summarization)
BERTScore	Semantic similarity	**0.72**	Low ($0.01/eval)	Quality filtering
METEOR	Stemming + synonyms	**0.55**	Free	Improved ROUGE
G-eval	LLM rates quality (multi-aspect)	**0.85**	Medium ($0.10/eval)	Detailed evaluation
GPT-4 Judge	Overall quality assessment	**0.92**	High ($0.30/eval)	Final validation
Faithfulness	Fact verification against source	**0.88**	High ($0.25/eval)	Critical applications
Human Eval	Gold standard (definition)	**1.00**	Very High ($5-20/eval)	Ground truth

Higher correlation with human judgment = better proxy for quality assessment

High Correlation (≥0.85) Medium Correlation (0.70-0.84) Low Correlation (<0.70)