Category	Metric	Definition /	Measurement Approach	Purpose
Automated	Perplexity	Measures h	Computed using log-likelihood on a held-	Evaluates model fluency and text quality.
Automated	BLEU (Bilingual Evaluation Unders	Compares g	Automated NLP metric (n-gram precision	Tests accuracy and content overlap.
Automated	ROUGE (Recall-Oriented Understu	Evaluates c	Uses lexical similarity and longest commc	Measures informativeness and completeness.
Automated	Semantic Similarity	Measures r	Cosine similarity using sentence transforr	Evaluates contextual relevance and semantic alignment.
Cultural	CSI Score (Cultural Sensitivity Inde	Weighted c	Calculated using human or model-assiste	Assesses overall cultural sensitivity.
Cultural	Stereotype Detection	Identifies p	Uses bias classifiers, keyword detection, (	Ensures inclusivity and avoids cultural generalizations.
Cultural	Regional Appropriateness	Evaluates v	Human evaluation or rule-based classifier	Tests localization and contextual fit.
Biblical Integration	Contextualization Accuracy	Measures v	Expert or rule-based evaluation comparir	Ensures meaningful and accurate cross-religious analogies.
Biblical Integration	Relevance	Evaluates h	Human scoring on relevance and interpre	Tests cross-cultural interpretive value.
Cost	Tokens Used	Number of	Retrieved from API logs or model output	Assesses resource usage.
Cost	API Costs	Monetary c	Based on token usage and provider pricin	Enables cost-performance comparison.
Cost	Inference Time	Average tin	Measured per API call or benchmark scrip	Evaluates computational efficiency.