Revisiting Metric Reliability for Fine-grained Evaluation of Machine Translation and Summarization in Indian Languages

Anonymous ACL submission

Abstract

While automatic metrics drive progress in Machine Translation (MT) and Text Summarization (TS), existing metrics have been developed and validated almost exclusively for English and other high-resource languages. This narrow focus leaves Indian languages—spoken by over 1.5 billion people—largely overlooked, casting doubt on the universality of current evaluation practices. To address this gap, we introduce **ITEM**, a large-scale benchmark that systematically evaluates the alignment of 26 automatic metrics with human judgments across six major Indian languages, enriched with fine-grained annotations. Our extensive evaluation—covering agreement with human judgments, sensitivity to outliers, language-specific reliability, inter-metric correlations, and resilience to controlled perturbations—reveals four central findings: (1) LLM-based evaluators show the strongest alignment with human judgments at both segment and system levels; (2) outliers exert a significant impact on metric-human agreement; (3) In TS, metrics are more effective at capturing content fidelity, whereas in MT, they better reflect fluency; and (4) Metrics differ in their robustness and sensitivity when subjected to diverse perturbations. Collectively, these findings offer critical guidance for advancing metric design and evaluation in Indian languages.[1]

1 Introduction

Effective deployment of Machine Translation (MT) and Text Summarization (TS) systems hinges on reliable evaluation. While human judgments remain the gold standard, their high cost, slow turnaround, and inherent subjectivity limit scalability and complicate large-scale or real-time assessment. Consequently, automatic metrics serve as practical proxies. Traditional surface-based metrics (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)) and modern embedding- or pretrained-based scorers (e.g., BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020)) offer more efficient alternatives. However, these metrics were originally developed and validated on datasets in high-resource languages such as English and German, raising questions about their reliability and generalizability to other languages, particularly low-resource Indian languages. At the same time, most research in MT and TS continues to rely on these metrics without empirical evidence of their efficacy (Datta et al., 2023; Urlana et al., 2023; Kumar et al., 2025).

Despite the rich linguistic diversity and widespread use of Indian languages, prior research has largely overlooked them in the evaluation of MT and TS systems, relying on a narrow set of metrics and lacking fine-grained analysis. Existing efforts, such as Sai B et al. (2023), have assessed metric reliability only for MT and primarily at the overall quality level, leaving other tasks and evaluation dimensions unexplored. In this work, we broaden the scope to both MT and TS, introducing four and two fine-grained evaluation criteria, respectively, across six major Indian languages.

Another limitation of prior work on metric evaluation lies in the limited consideration of robustness, which we address in this paper along two dimensions. First, previous studies often ignore the impact of outliers. As noted by Mathur et al. (2020), correlation measures such as Pearson's can produce spurious correlations when outliers are present, yet works in Indian languages like Sai B et al. (2023) overlook this issue. Second, most recent evaluations rely on high-quality translations or summaries, which weakens robustness analysis by offering only shallow signals—such as token overlap or embedding similarity—that do not always align with human perception. In the context of Indian languages, assessing metric robustness is particularly important, especially under controlled

[1]This dataset is publicly released under a Creative Commons license at anon.com

perturbations such as paraphrasing, negation, syntactic reordering, and lexical substitutions (e.g., synonyms or antonyms). Given their flexible word order and a wide range of morphological complexity—from moderate to highly inflectional—Indian languages exemplify the challenges of robust metric evaluation (Tandon and Sharma, 2017).

Our contributions can be summarized as follows:

- We introduce **ITEM** (**I**ndian **T**ext **E**valuation **M**etrics Testbed), a benchmark for assessing the alignment between automatic metrics and human judgments in MT and TS across six major Indian languages: Hindi, Bengali, Marathi, and Gujarati (Indo-Aryan), and Tamil and Telugu (Dravidian).
- We conduct a comprehensive evaluation of 26 automatic metrics to identify those that best align with human judgments at both segment and system levels.
- We examine how effectively existing metrics capture diverse evaluation dimensions, including adequacy, fluency, coherence, and factuality.
- We investigate the effect of outlier samples on metric–human correlations and evaluate metric robustness and sensitivity under controlled perturbations such as paraphrasing, negation, and syntactic reordering.

2 Related Work

Evaluation of Machine Translation System BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), COMET (Rei et al., 2020), and LLM-based scores (Yuan et al., 2021; Fu et al., 2024) are among the most commonly used metrics for evaluating MT. However, these evaluations have generally been limited in scope, focusing on a narrow range of languages and metrics. Although some recent studies (Wu et al., 2024; Agrawal et al., 2024; Freitag et al., 2024) provide broader quality assessments, they entirely exclude Indian languages. Existing works that do include Indian languages (Sai B et al., 2023; Singh et al., 2024) remain incomplete, as they report only overall scores, overlook multiple evaluation dimensions, and cover fewer languages than our study.

Evaluation of Text Summarization Systems The most commonly used metrics for evaluating TS are ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and LLM-based scores (Fu et al., 2024). In contrast, older metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), as well as newer ones like MoverScore (Zhao et al., 2019), are less frequently adopted. The scope of TS evaluation has also expanded beyond recall-based coverage to include additional dimensions such as focus, faithfulness, and inter-sentential coherence (Koto et al., 2022; Fabbri et al., 2021).

Existing studies, however, have primarily focused on English, with only limited multilingual efforts (Koto et al., 2021; Clark et al., 2023; Han et al., 2024) extending coverage to other languages. These works still exclude Indian languages and rely on a narrow set of traditional metrics and evaluation aspects.

Robustness Studies on Evaluation Systems The robustness of evaluation metrics under perturbations has received limited attention. Guo et al. (2024) investigates controlled perturbations for English Plain Language Summarization.[2] Huang and Baldwin (2023) evaluates MT metrics under word- and character-level attacks, yet excludes TS, Indian languages, and a broader range of perturbations. Other works (Alves et al., 2022; Chen et al., 2022) focus solely on MT, while (Patel et al., 2024) considers only English TS. None provides a comprehensive robustness assessment of both TS and MT metrics across Indian languages.

In contrast to prior work, we present a unified, high-resolution evaluation framework for both TS and MT in Indian languages. Our study goes beyond single overall scores by assessing multiple quality dimensions using 26 diverse metrics across varied configurations (including 48 ROUGE, 144 BERTScore, and 4 BLEU variants). We further evaluate robustness through controlled paraphrasing and perturbation tests, providing a benchmark-level, fine-grained analysis that surpasses the scope of existing studies.

3 ITEM

3.1 Dataset Construction

The **ITEM** benchmark was developed by assembling parallel datasets for summarization and translation in six widely spoken Indian languages: Hindi, Bengali, Tamil, Telugu, Marathi, and Gujarati. For summarization, we randomly sampled 150 article–summary pairs per language from the XLSum cor-

[2] A task of generating semantically faithful yet structurally simplified summaries to improve accessibility for readers with limited linguistic or domain expertise.

Language	Worker 1		Worker 2	
	Sum.	Trans.	Sum.	Trans.
Hindi	81.25	92.5	86.25	85
Bengali	80	77.5	77.5	85
Tamil	85	82.5	76.25	82.5
Telugu	92.5	87.5	87.5	87.5
Marathi	76.25	77.5	82.5	87.5
Gujarati	87.5	87.5	91.25	90

Table 1: Annotators quality (%).

pus (Hasan et al., 2021), which provides professionally annotated summaries of BBC news articles. For translation, we randomly selected 150 sentence–translation pairs per language from FLORES-200 (Team et al., 2022), a widely used benchmark with human-verified translations.

To obtain a broad spectrum of machine-generated outputs, we applied three complementary models to the sampled data. The set comprises two state-of-the-art proprietary LLMs—Cohere Command R+ (Cohere Labs, 2024) and GPT-4o mini (OpenAI et al., 2024)—alongside IndicBART and IndicTrans2 from AI4Bharat (Dabre et al., 2022; Gala et al., 2023), open-source models designed with a focus on Indian languages. Please refer to Figure 1 for the complete annotation process **ITEM**.

3.2 Evaluation Aspects

To account for the multifaceted nature of MT and TS quality, **ITEM** employs a set of carefully adopted quality indicators. For MT, evaluation centers on two fundamental dimensions (Koehn and Monz, 2006): (1) *Adequacy*, which measures how accurately the translation conveys the meaning and intent of the source without omissions or distortions, and (2) *Fluency*, which assesses the readability of the translated text in the target language, focusing on grammaticality, natural expression, and stylistic appropriateness.

For TS, a more nuanced four-dimensional framework is adopted to reflect the task's inherent complexity (Koto et al., 2022). (1) *Faithfulness* measures factual consistency, ensuring the summary avoids hallucinations or contradictions. (2) *Focus* assesses the precision of the summary, ensuring only relevant information is included. (3) *Coverage* examines how well the summary captures the key and salient ideas of the source. (4) *Coherency* measures the logical continuity between sentences, ensuring that the summary maintains a seamless flow and remains readily understandable.

3.3 Quality Control

Reliability of human annotations was a cornerstone in developing **ITEM**. To ensure consistency, we first established detailed annotation guidelines and recruited **two** native speakers per target language to evaluate the extracted samples. All annotators were qualified experts, each holding at least a high school diploma and proficient in both English and the respective Indian language. Annotators scored each sample against the predefined quality indicators using a five-point scale (1 = Poor, 5 = Excellent), with every level explicitly defined to minimize ambiguity and reduce subjective variance.

Beyond initial annotation, we implemented a structured quality control (QC) procedure. For each task and language, 20 representative samples[3] were selected and assessed by five independent expert annotators[4], who provided both scores and written justifications. These QC annotations were carefully reviewed to verify alignment and reliability. Final reference scores for each of the 20 QC samples were obtained by averaging the ratings from all five annotators.

The reliability of our two main workers was then evaluated against these 20 QC samples. Specifically, we computed the proportion of annotations by each worker that deviated by at most one point from the QC reference scores across all evaluation aspects. This metric served as a robust and interpretable measure of annotator quality. The resulting quality scores are summarized in Table 1.

3.4 Data Statistics

The initial release of **ITEM** included 5,400 samples (2 tasks $\times$ 150 datapoints $\times$ 3 models $\times$ 6 languages). To enhance the benchmark's reliability, we filtered out samples with high annotator disagreement, resulting in 2,571 samples for summarization and 2,604 for translation, prioritizing annotation consistency over sheer volume. Table 2 reports inter-annotator correlations post-refinement, capturing alignment between annotators.

Tables 3 and 4 present the average token counts for source texts alongside human references, as well as machine-generated summaries and transla-

[3]These 20 samples, encompassing both high- and low-quality data, were strategically selected to rigorously assess annotator sensitivity across the quality spectrum and strengthen QC reliability.

[4]All annotators in this study were aged 17–22, native speakers of their respective language, and proficient in English. Each worker is fairly compensated in accordance with an agreement, with payments aligned to India's minimum wage.

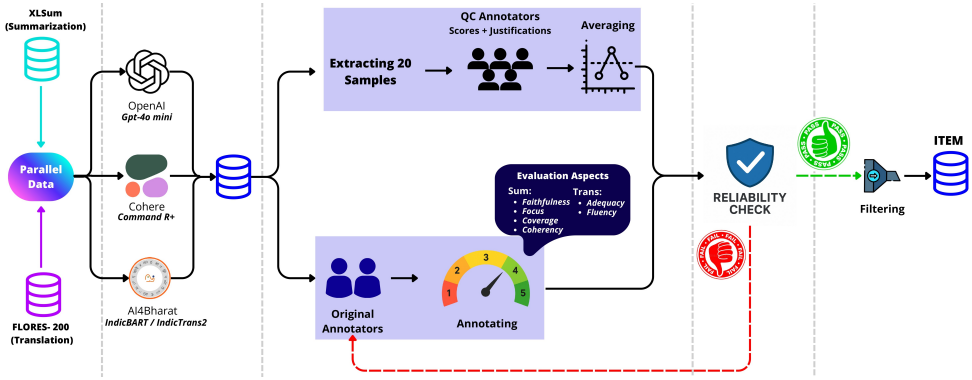


Figure 1: End-to-End process of dataset creation.

Language	Summarization					Translation		
	Fa.	Fo.	Cov.	Coh.	Avg	Ad.	Fl.	Avg
Hindi	0.53	0.53	0.48	0.55	0.74	0.54	0.63	0.61
Bengali	0.8	0.82	0.84	0.4	0.93	0.54	0.47	0.64
Tamil	0.45	0.57	0.68	0.71	0.81	0.65	0.61	0.72
Telugu	0.74	0.69	0.72	0.73	0.89	0.78	0.76	0.87
Marathi	0.58	0.57	0.55	0.58	0.79	0.59	0.49	0.44
Gujarati	0.85	0.83	0.86	0.65	0.91	0.75	0.75	0.77

Table 2: Inter-annotator Pearson correlations across evaluation aspects: Faithfulness (Fa), Focus (Fo), Coverage (Cov), Coherence (Coh), Adequacy (Ad), and Fluency (Fl).

Language	Article	Summary	English Text	Translation
Hindi	2,476	145	139	136
Bengali	3,344	151	139	135
Tamil	3,516	195	139	161
Telugu	3,930	170	139	138
Marathi	3,601	176	139	142
Gujarati	3,259	136	139	133

Table 3: Average token counts for source texts and human-generated summaries/translations.

tions, providing a concise overview of text length variations across different tasks and languages.

We define human scores as the average of the ratings provided by two annotators. Figure 2 depicts the distribution of human evaluation scores. In TS, Bengali attains the highest quality ratings, while Hindi ranks lowest. Conversely, in MT, Hindi leads in quality, with Marathi at the bottom. A deeper examination of these scores reveals that GPT-4o mini consistently excels in TS across all languages, whereas IndicTrans2 dominates MT for most languages, with GPT-4o mini outperforming only in Hindi.

From Figure 2, we observe that the distributions of *Faithfulness*, *Focus*, and *Coverage* are largely similar. Consistently, Figure 3 highlights their strong inter-metric correlations, whereas *Coherency* exhibits notably weaker association. Meanwhile, *Adequacy* and *Fluency* show limited mutual dependence, reflecting their distinct evaluative roles.

Language	Cohere		GPT		AI4Bharat	
	Sum.	Trans.	Sum.	Trans.	Sum.	Trans.
Hindi	239	140	268	136	194	142
Bengali	138	135	248	136	228	136
Tamil	120	162	245	157	245	158
Telugu	82	143	222	138	228	137
Marathi	149	138	241	137	231	139
Gujarati	84	140	223	132	183	136

Table 4: Average token counts for machine-generated summaries/translations.

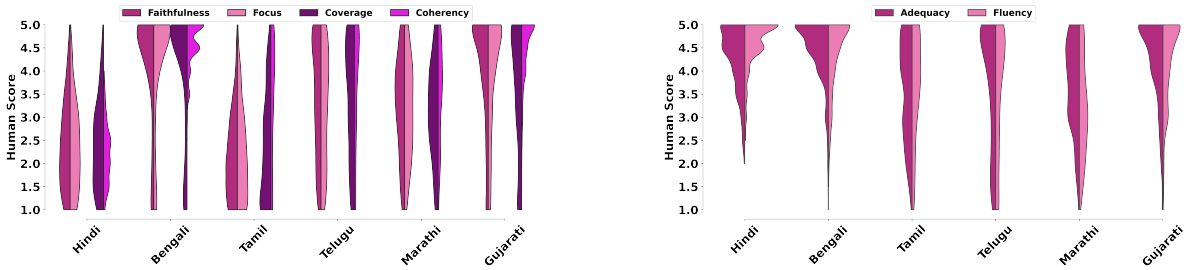
4 Experiments

4.1 Evaluation Metrics

To achieve a holistic and nuanced assessment, we employ a diverse collection of metrics that span surface-level text similarity, semantic understanding, learned human-aligned judgments, and cutting-edge LLM-based evaluation.

Lexical and n-gram Overlap Metrics. These metrics quantify direct textual correspondences between hypotheses and references. Our set includes BLEU (1-4) (Papineni et al., 2002), sentBLEU (Chen and Cherry, 2014), SacreBLEU (Post, 2018), GLEU (Mutton et al., 2007), NIST (Doddington, 2002), LEPOR (Han et al., 2012), ChrF (Popović, 2015), ChrF++ (Popović, 2017), METEOR (Banerjee and Lavie, 2005), and a comprehensive suite of ROUGE variants (Lin, 2004). Specifically, we employ eight ROUGE variants—ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-W, ROUGE-S4, and ROUGE-SU4. Each variant is evaluated under two stopword settings (with/without stopwords) and three scoring modes (precision, recall, F1-score), resulting in 48 distinct configurations.

Embedding-based Semantic Metrics. To capture semantic alignment beyond surface matches, we utilize embedding-based approaches, including BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), MoverScore (Zhao et al., 2019), Embedding Averaging (Landauer and Du-



(a) Text Summarization (b) Machine Translation

Figure 2: Distribution of human evaluation scores across evaluation aspects.

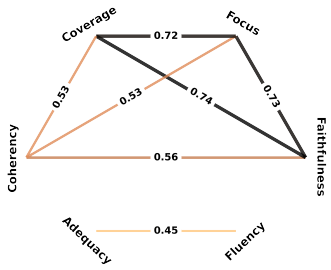


Figure 3: Pearson correlation network of human evaluation aspects.

mais, 1997), Greedy Matching (Rus and Lintean, 2012), Vector Extrema (Forgues et al., 2014), LaBSE (Feng et al., 2022), and LASER (Artetxe and Schwenk, 2019). BERTScore is computed using four models: IndicBERT (Kakwani et al., 2020), mBERT cased, mBERT uncased (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020), across all 12 layers and using precision, recall, and F-score, yielding 144 distinct configurations.

Neural Learned Metrics. Metrics specifically trained to emulate human judgment include BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020), offering refined evaluation of nuanced language understanding.

LLM-based Evaluation. At the forefront, we harness the reasoning and discourse-aware capabilities of large language models for evaluation. GPTScore (Fu et al., 2024), implemented with Flan-T5-XXL, evaluates outputs along six predefined dimensions, providing diagnostic insights that extend beyond overall scores. Complementing this, we utilize state-of-the-art LLMs—including GPT-4.1, LLaMA-3.3 (Grattafiori et al., 2024), DeepSeek V3 (DeepSeek-AI et al., 2025), and Gemini 2.5 (Comanici et al., 2025)—through designed prompting protocols (full prompt details in Appendix A.4).

Full descriptions and implementation details for all metrics are available in Appendix A.

4.2 Outliers Detection

We characterize outliers as observations that deviate markedly from the rest of the dataset (Barnett et al., 1994). Given Pearson correlation's sensitivity to such anomalies, we applied a robust detection method on the human scores for each evaluation aspect. Rather than relying on standardization, which is vulnerable to extreme values due to its dependence on the mean and standard deviation, we employed the median and Median Absolute Deviation (MAD) for greater resilience (Iglewicz and Hoaglin, 1993; Leys et al., 2013). For a set of human scores s, MAD was computed as:

$$\text{MAD} = 1.483 \times \text{median}(|s - \text{median}(s)|)$$

A robust z-score was then derived using:

$$z = \frac{s - \text{median}(s)}{\text{MAD}}$$

Observations with $|z| > 3.5$ were flagged as outliers (Mathur et al., 2020). Table 5 reports the counts of detected outliers across languages and indicators.

Language	Summarization				Translation	
	Fa.	Fo.	Cov.	Coh.	Ad.	Fl.
Hindi	0	0	0	0	0	0
Bengali	44	50	61	0	8	8
Tamil	2	5	0	0	0	0
Telugu	0	0	0	0	0	0
Marathi	0	0	0	0	0	3
Gujarati	37	45	0	0	20	22

Table 5: Outlier distribution across languages for evaluation aspects: Faithfulness (Fa), Focus (Fo), Coverage (Cov), Coherence (Coh), Adequacy (Ad), and Fluency (Fl).

4.3 Results and Analysis

4.3.1 Segment-level Analysis

Table 6 presents our segment-level analysis, highlighting Pearson correlations between automatic metrics and human judgments.

We identified the optimal configurations for

Metric	Summarization				Translation		Overall
	Faithfulness	Focus	Coverage	Coherency	Adequacy	Fluency	
BLEU-3	0.125 / 0.114	0.161 / 0.142	0.156 / 0.163	0.087 / 0.087	0.21 / 0.189	0.276 / 0.265	0.169 / 0.16
sentBLEU	0.121 / 0.115	0.153 / 0.138	0.145 / 0.156	0.093 / 0.093	0.209 / 0.187	0.275 / 0.264	0.166 / 0.159
SacreBLEU	0.074 / 0.051	0.09 / 0.064	0.068 / 0.058	0.013 / 0.013	0.138 / 0.132	0.192 / 0.184	0.096 / 0.084
GLEU	0.079 / 0.078	0.113 / 0.099	0.103 / 0.114	0.085 / 0.085	0.206 / 0.184	0.27 / 0.259	0.143 / 0.137
NIST	0.098 / 0.124	0.105 / 0.114	0.104 / 0.128	0.105 / 0.105	0.143 / 0.126	0.191 / 0.185	0.124 / 0.13
METEOR	0.115 / 0.098	0.138 / 0.122	0.141 / 0.134	0.025 / 0.025	0.169 / 0.159	0.239 / 0.231	0.138 / 0.128
LEPOR	0.187 / 0.175	0.167 / 0.166	0.19 / 0.195	0.075 / 0.075	0.1 / 0.08	0.095 / 0.096	0.136 / 0.131
ROUGE-2-N-R	0.258 / **0.258**	0.248 / 0.249	**0.279 / 0.294**	0.141 / 0.141	0.189 / 0.165	0.246 / 0.241	0.227 / 0.225
ChrF	0.222 / 0.214	0.24 / 0.23	0.256 / 0.265	0.12 / 0.12	0.2 / 0.18	0.266 / 0.255	0.217 / 0.211
ChrF++	0.205 / 0.191	0.23 / 0.214	0.239 / 0.245	0.116 / 0.116	0.201 / 0.181	0.269 / 0.257	0.21 / 0.201
BERTScore	0.208 / 0.198	0.239 / 0.233	0.251 / 0.259	0.119 / 0.119	0.244 / 0.222	0.31 / 0.303	0.229 / 0.222
BARTScore	0.106 / 0.13	0.121 / 0.134	0.119 / 0.132	0.05 / 0.05	0.058 / 0.068	0.108 / 0.11	0.094 / 0.104
MoverScore	0.09 / 0.093	0.094 / 0.098	0.109 / 0.065	0.094 / 0.094	0.059 / 0.068	0.107 / 0.117	0.092 / 0.089
Embedding Averaging	0.12 / 0.1	0.196 / 0.175	0.152 / 0.151	0.047 / 0.047	0.259 / 0.237	0.305 / 0.297	0.18 / 0.168
Greedy Matching	0.121 / 0.101	0.199 / 0.177	0.152 / 0.151	0.048 / 0.048	0.259 / 0.236	0.307 / 0.298	0.181 / 0.169
Vector Extrema	0.121 / 0.101	0.199 / 0.177	0.152 / 0.151	0.048 / 0.048	0.259 / 0.236	0.307 / 0.298	0.181 / 0.169
LaBSE	0.132 / 0.112	0.185 / 0.169	0.172 / 0.172	0.081 / 0.081	0.286 / 0.264	0.346 / 0.34	0.2 / 0.19
LASER	0.048 / 0.086	0.104 / 0.129	0.086 / 0.079	0.06 / 0.06	0.093 / 0.114	0.132 / 0.147	0.087 / 0.103
BLEURT	0.175 / 0.188	0.201 / 0.208	0.223 / 0.226	0.163 / 0.163	0.291 / **0.289**	0.362 / 0.361	0.236 / 0.239
COMET	0.214 / 0.24	0.24 / 0.262	0.251 / 0.27	0.204 / 0.204	0.267 / 0.255	0.339 / 0.339	0.253 / 0.262
GPTScore	-0.053 / -0.044	-0.074 / -0.048	-0.062 / -0.047	-0.048 / -0.048	0.008 / -0.011	-0.003 / -0.009	-0.039 / -0.035
LLama-3.3-Instruct	0.228 / 0.219	0.252 / 0.242	0.24 / 0.244	0.159 / 0.159	0.263 / 0.232	0.335 / 0.334	0.246 / 0.238
DeepSeek-V3	0.251 / 0.234	**0.289** / 0.269	0.277 / 0.275	0.233 / 0.233	**0.311** / 0.282	**0.379 / 0.364**	**0.29 / 0.276**
GPT-4.1 (English prompt)	**0.263** / 0.257	0.278 / 0.271	0.238 / 0.244	**0.24 / 0.24**	0.274 / 0.241	0.354 / 0.343	0.275 / 0.266
GPT-4.1 (Native prompt)	0.235 / 0.242	0.241 / 0.254	0.245 / 0.254	0.168 / 0.168	0.289 / 0.262	0.359 / 0.347	0.256 / 0.255
Gemini 2.5 Flash	0.248 / 0.244	0.277 / **0.274**	0.257 / 0.264	0.202 / 0.202	0.295 / 0.265	0.363 / 0.347	0.274 / 0.266

Table 6: Segment-level Pearson correlations between automatic metrics and human judgments; values after '/' reflect results with outliers removed.

three key metrics based on segment-level correlations with human assessments. Out of 48 ROUGE configurations explored, ROUGE-2 with recall scoring and stopwords removed emerged as the top performer. For BERTScore, IndicBERT at layer 8 with recall mode yielded the best results, while BLEU-3 was the most effective BLEU configuration. All subsequent references to these metrics adopt these high-performing settings.

Performance trends varied across metric categories. Lexical overlap metrics showed task-specific strengths: ROUGE excelled in TS, whereas ChrF++ led in MT. Among embedding-based metrics, BERTScore was strongest for TS, while LaBSE dominated in MT. Neural learned metrics exhibited similar task preferences, with COMET outperforming in TS and BLEURT aligning better with MT human judgments.

LLM-based evaluators, however, displayed a consistent dominance. DeepSeek-V3 achieved the highest correlations for both TS and MT, establishing itself as the most reliable metric overall. Prompting also influenced performance: native-language prompts outperformed English in MT, whereas English prompts generally fared better in TS, except for *Coverage*. Overall, LLM-based and neural learned metrics proved more dependable at the segment level than traditional metrics. A detailed comparison at the system level is provided in Appendix C.1.

A complementary pattern emerges between summarization and translation evaluation. In TS, automatic metrics align closely with content-oriented dimensions like *Coverage* and *Focus*, but are less sensitive to *Coherency*, highlighting limitations in capturing logical flow and overall readability. Conversely, MT evaluation prioritizes *Fluency*, with *Adequacy* reasonably captured, while finer semantic and discourse properties remain less reliably assessed. This suggests summarization metrics better track content preservation, whereas translation metrics better reflect how content is expressed, emphasizing a trade-off between content fidelity and linguistic quality.

Outliers significantly affect human–metric correlations, though their impact differs across indicators. In TS, removing outliers generally reduces correlations for *Faithfulness* and *Focus*, indicating previous inflation due to extreme samples. LASER, however, shows a substantial increase (+79.17% for *Faithfulness*, +24.04% for *Focus*[5]), suggesting prior underestimation. In contrast, *Coverage* typically improves across most metrics following outlier removal, with a few exceptions—such as MoverScore and SacreBLEU. *Coherency* remains unaffected, consistent with the absence of outliers reported in Table 5.

For MT, excluding outliers generally decreases correlations for both *Adequacy* and *Fluency*, except

[5]Percent changes quantify outlier effects as the relative difference between correlations with and without outliers, expressed as a percentage of the original correlation.

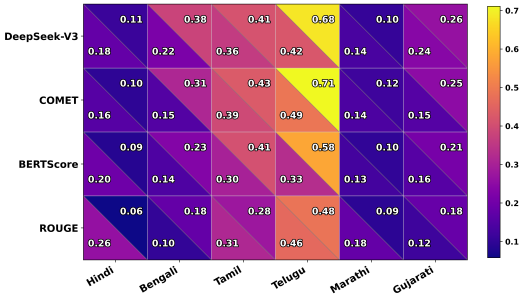


Figure 4: Language-specific Pearson correlations of top metrics across tasks (top-right: MT, bottom-left: TS).

for metrics like LASER, BARTScore, and MoverScore, which show improved alignment with human judgments.

Sensitivity to outliers varies widely across metrics. LASER (+18.39%), BARTScore (+10.64%), and SacreBLEU (-12.5%) exhibit substantial shifts, whereas robust metrics such as ROUGE (-0.88%) and GPT-4.1(Native prompt) (-0.39%) remain stable. Overall, embedding-based metrics are more vulnerable to outliers, while neural and LLM-based evaluators demonstrate consistent resilience.

Crucially, to our knowledge, this study is the first to systematically investigate the influence of outliers on human–metric correlations within Indian languages. Previous research in this area has not addressed this issue, underscoring both the novelty and the importance of our findings.

4.3.2 Language Effects

Figure 4 illustrates language-specific Pearson correlations for the top-performing metrics across both tasks, highlighting clear variations in performance across languages. Telugu consistently achieves the highest scores, whereas Marathi in TS and Hindi in MT rank lowest. Interestingly, LLM-based evaluations follow the same pattern. Notably, Hindi—a high-resource language—is evaluated less reliably than lower-resource languages like Tamil and Telugu, revealing an unexpected anomaly in LLM-based assessments that contrasts with previous evidence of superior performance in high-resource settings (Li et al., 2025; Yari and Koto, 2025).

4.3.3 Metric Inter-correlation

Figures 5 present the Pearson correlation matrices of automatic metrics, uncovering well-defined clusters that reflect our proposed taxonomy. Lexical overlap metrics formed a compact, highly correlated group, reaffirming their common reliance on surface-level n-gram matching. Embedding-based metrics likewise emerged as a cohesive cluster, while vector-based metrics displayed perfect internal agreement.

LLM-based evaluators distinguished themselves with strikingly high mutual correlations ($r \approx 0.7-0.9$) across both tasks, underscoring their consistent judgment patterns despite architectural diversity. Yet, their only moderate alignment with traditional metrics underscores a methodological divide between evaluation grounded in large language models and established paradigms.

Notably, LEPOR, MoverScore, and LASER persistently showed the weakest average correlations with other metrics, indicating either that they capture unique quality dimensions or diverge from the prevailing consensus of evaluation methods.

4.4 Robustness & Sensitivity

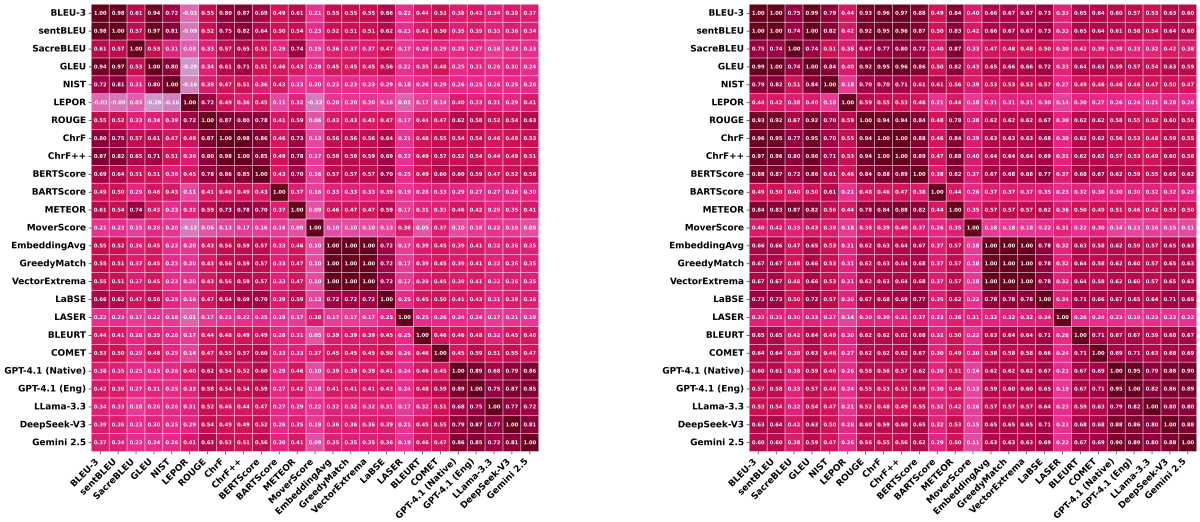
4.4.1 Paraphrasing Sensitivity

To evaluate the robustness of automatic metrics beyond surface word overlap, we conducted a controlled paraphrasing experiment. GPT-4.1 rephrased human-written summaries (prompt in Appendix B), followed by a two-step validation. In each language, two native speakers assessed the outputs based on four criteria: meaning preservation, grammaticality, clarity, and detail consistency. A primary annotator first reviewed and minimally corrected borderline cases, after which a second annotator verified quality, retaining only paraphrases with unanimous approval. Figure 6 shows the number of validated paraphrases out of 150 per language.

4.4.2 Noise Perturbations

We evaluated metric resilience under controlled perturbations affecting semantic fidelity, surface fluency, and summarization-specific traits. By introducing noise to reference outputs, we simulated real-world imperfections to assess each metric's reliability and sensitivity across linguistic dimensions.

Semantic Perturbations focused on meaning alterations. Using IndicNER—a multilingual named entity recognition model for Indian languages (Mhaske et al., 2023)—we constructed lexicons of persons, locations, and organizations. For each text, a randomly selected entity was substituted, deleted, corrupted at the character level, or augmented with an additional entity of the same type. Further semantic modifications included numeric adjustments, verb negation using language-specific markers, and synonym or antonym replacements.



(a) Text Summarization

(b) Machine Translation

Figure 5: Pearson correlation matrices of automatic metrics.

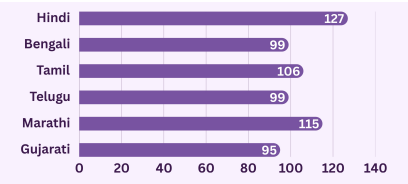


Figure 6: Distribution of high-quality paraphrased summaries across languages.

These manipulations primarily affect MT *Adequacy* and TS *Faithfulness*.

Lexical and Structural Perturbations targeted MT *Fluency* and TS *Coherency*. In translations, words were shuffled within a four-token window, whereas summaries experienced sentence-level shuffling through pairwise swaps. Additional noise included random stopword insertions or deletions, as well as character-level edits[6].

Summarization-specific Perturbations probed higher-level content properties. To evaluate *Focus*, we added a contextually plausible but off-topic sentence from another summary. To challenge *Coverage*, we removed the sentence with the lowest ROUGE overlap, simulating the loss of the most informative content.

To describe the results, we focus on COMET as one of the most reliable open-source evaluation metrics, with complementary analyses for ROUGE and BERTScore provided in Appendix C.4. Figure 13 shows COMET's relative score drops across languages under different perturbations. In MT, shuffling is the most disruptive, while in TS, truncating key sentences leads to the greatest decline. Paraphrasing and synonym replacements have minimal impact, indicating COMET's robustness to meaning-preserving variations. Negation effects vary by language: Tamil and Telugu are highly sensitive, while others remain stable. Overall, Hindi is most vulnerable and Gujarati most resilient.

Figure 14 presents the relative declines in human–COMET correlations across evaluation aspects. In MT, shuffling causes the sharpest drops in *Adequacy* and *Fluency*; in TS, truncation reduces correlations for *Faithfulness*, *Focus*, and *Coverage*, but unexpectedly increases *Coherency*. Overall, *Focus* remains the most stable aspect, whereas *Fluency* is least robust. These findings highlight COMET's strengths in handling lexical variation but also its limitations under structural and meaning-altering perturbations, reinforcing the importance of targeted robustness evaluations across languages and quality dimensions.

5 Conclusion

This study introduces **ITEM**, a benchmark for evaluating automatic metrics in MT and TS across six Indian languages and quality dimensions, including alignment with human judgments, outlier effects, language consistency, inter-metric correlations, and robustness under perturbations. Results show that LLM-based evaluators correlate most strongly with human ratings, with metrics emphasizing fidelity in TS and fluency in MT. Robustness analyses reveal vulnerabilities to structural and semantic noise, highlighting the need for more resilient evaluation methods. Overall, **ITEM** provides a foundation for developing future metrics that are reliable and interpretable for Indian languages.

[6]replacement, insertion, deletion, or transposition

Limitations

Despite offering a comprehensive framework for evaluating MT and TS metrics in Indian languages, this study is not without limitations. Our analysis is confined to six widely spoken languages, leaving out many low-resource Indian languages whose rich typological variation could divulge further vulnerabilities in existing metrics. Moreover, while LLM-based evaluators showed the strongest alignment with human judgments, their reliance on resource-heavy, opaque, and often proprietary systems raises concerns about accessibility, reproducibility, and interpretability—highlighting the need for developing transparent and efficient alternatives. Finally, although our robustness experiments covered a diverse set of controlled perturbations, they cannot fully capture the complexity of real-world textual noise.

Future work should explore hybrid strategies that combine human-in-the-loop feedback, linguistic knowledge, and LLM reasoning to build more reliable and inclusive multilingual evaluation frameworks.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Vic Barnett, Toby Lewis, and 1 others. 1994. *Outliers in statistical data*, volume 3. Wiley New York.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Cohere Labs. 2024. c4ai-command-r-plus-08-2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing, pages 5291–5302, Singapore. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Preprint*, arXiv:2007.12626.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Rilyn Han, Jiawen Chen, Yixin Liu, and Arman Cohan. 2024. Rethinking efficient multilingual text summarization meta-evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15739–15746, Bangkok, Thailand. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703. Association for Computational Linguistics.

Yichen Huang and Timothy Baldwin. 2023. Robustness tests for automatic machine translation metrics with adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5126–5135, Singapore. Association for Computational Linguistics.

Boris Iglewicz and David C Hoaglin. 1993. *Volume 16: how to detect and handle outliers*. Quality Press.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian

languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. FFCI: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Raghvendra Kumar, Mohammed Salman S A, Aryan Sahu, Tridib Nandi, Pragathi Y P, Sriparna Saha, and Jose G Moreno. 2025. COSMMIC: Comment-sensitive multimodal multilingual Indian corpus for summarization and headline generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8728–8748, Vienna, Austria. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying LLM performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kevin Patel, Suraj Agrawal, and Ayush Kumar. 2024. Tweak to trust: Assessing the reliability of summarization metrics in contact centers via perturbed summaries. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 172–186, Mexico City, Mexico. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.

Juhi Tandon and Dipti Misra Sharma. 2017. Unity in diversity: A unified parsing strategy for major Indian languages. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 255–265, Pisa, Italy. Linköping University Electronic Press.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, and Barry Haddow. 2023. PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11606–11628, Singapore. Association for Computational Linguistics.

Guojun Wu, Shay B Cohen, and Rico Sennrich. 2024. Evaluating automatic metrics with incremental machine translation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2994–3005, Miami, Florida, USA. Association for Computational Linguistics.

Amir Hossein Yari and Fajri Koto. 2025. Unveiling cultural blind spots: Analyzing the limitations of mLLMs in procedural text comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20151–20170, Vienna, Austria. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Evaluation Metrics

A.1 Lexical and n-gram Overlap Metrics

BLEU: The Bilingual Evaluation Understudy (BLEU) quantifies the similarity between a generated hypothesis and a reference translation. It captures adequacy through word-level precision and fluency via n-gram precision across n = 1 to 4. BLEU computation integrates three components: (1) n-gram precision comparing the output with reference translations, (2) a brevity penalty (BP) to avoid inflated scores from overly short sentences, and (3) clipping to regulate repeated word occurrences.

SentBLEU: A sentence-level adaptation of the BLEU metric, evaluates each sentence independently rather than averaging n-gram precisions across an entire corpus, offering a more granular assessment of machine-generated quality.

SacreBLEU: SacreBLEU provides a standardized, reproducible variant of the BLEU metric, ensuring consistent evaluation across studies. By enforcing uniform tokenization, normalization, and punctuation handling, it mitigates discrepancies arising from differing preprocessing or parameter choices in traditional BLEU implementations.

GLEU: Google BLEU (GLEU) adapts BLEU for sentence-level evaluation, making it ideal for assessing individual machine-generated outputs. It measures the overlap of n-grams between a candidate and a reference sentence, while applying a length penalty to discourage overgeneration, providing a more precise reflection of sentence-level quality.

NIST: The metric, developed by the National Institute of Standards and Technology, evaluates similarity by assigning weights to matching n-grams according to their informational content. Rare n-gram sequences receive higher weights, reflecting their greater contribution to conveying meaningful content compared to more common sequences.

LEPOR: LEPOR (Length Penalty, Precision, n-gram Position Difference Penalty, and Recall) is a robust, language-independent metric. It integrates multiple factors—an enhanced length penalty, precision, n-gram position difference penalty, and recall—without relying on external linguistic resources.

ROUGE Metrics: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) suite quantifies the overlap between a machine-generated text and a reference, focusing on shared textual units such as n-grams or word sequences. Key variants include ROUGE-N (n-gram overlap), ROUGE-L (longest common subsequence), ROUGE-W (weighted LCS), ROUGE-S4 (skip-bigrams within a fixed window), and ROUGE-SU4 (skip-bigrams plus unigrams). Each metric offers precision, recall, and F-score measures, enabling a comprehensive assessment of summary quality.

ChrF / ChrF++: The Character n-gram F-score (ChrF) evaluates the similarity between generated and reference texts at the character level, rather than using word n-grams. This approach reduces sensitivity to tokenization, making it robust across languages and text processing variations.

METEOR: METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates machine-generated text by combining a weighted F-score with a penalty function that balances precision and recall. The method begins by aligning the hypothesis with the reference to identify the longest matching sequences of words, incorporating synonyms from thesauri or large corpora to capture semantic equivalence and morphological variations. Precision and recall are computed based on matched unigrams, while the penalty function discourages short, fragmented matches and rewards longer, contiguous sequences. This design ensures a nuanced assessment of quality, emphasizing both accuracy and fluency.

A.2 Embedding-based Semantic Metrics

BERTScore: Evaluates the similarity between a hypothesis and reference by leveraging contextual embeddings from a pre-trained BERT model, which is trained for masked language modeling and next sentence prediction. It computes the cosine similarity between the token embeddings of the machine-generated output and the reference, using greedy matching to maximize alignment. The metric reports precision, recall, and F1-score, providing a nuanced assessment of semantic overlap beyond surface-level matching.

BARTScore: A text generation evaluation metric that leverages BART's encoder–decoder architecture to bridge the gap between training objectives and sentence generation. Trained with a denoising objective, BART is particularly effective for assessing conditional generation tasks. BARTScore computes the weighted log-probability of the reference sentence given the hypothesis, using the pre-trained parameters of BART, providing a precise measure of generation quality.

MoverScore: Quantifies the semantic similarity between a summary and its reference by applying the Word Mover's Distance on n-gram embeddings derived from BERT representations.

Embedding Averaging: Represents a sentence by averaging the embeddings of all its words and measures sentence similarity using cosine similarity.

Greedy Matching: Evaluates semantic alignment by pairing each word in the generated sentence with its closest semantic counterpart in the reference, producing an overall similarity score.

Vector Extrema: Constructs sentence embeddings by taking the dimension-wise maxima or minima of word embeddings, emphasizing dominant semantic features. The cosine similarity between these vectors quantifies the semantic similarity between the generated and reference text.

LaBSE & LASER: Language-agnostic BERT Sentence Embeddings (LaBSE) and Language-Agnostic SEntence Representations (LASER) are pre-trained multilingual models designed to produce dense, language-independent sentence embeddings. These embeddings allow direct semantic comparison of sentences across diverse languages. LaBSE leverages a BERT-based architecture tailored for cross-lingual tasks, whereas LASER uses a BiLSTM encoder trained on parallel corpora to achieve its multilingual representations.

A.3 Neural Learned Metrics

BLEURT: BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) is a BERT-based, pre-trained model that leverages multi-task loss on extensive synthetic datasets comprising numerous reference sentences. Designed as a sentence-level evaluation metric, it produces prediction scores that quantify the similarity between a generated hypothesis and reference texts, effectively approximating human judgment. Its core objective is to train regression models capable of predicting human evaluation scores accurately.

COMET: COMET (Cross-lingual Optimized Metric for Evaluation of Translation) is an evaluation framework that integrates ranking and regression techniques. It comprises an estimator and a reference-based ranking model grounded in human judgments. Employing a dual cross-lingual modeling approach for encoder inputs, COMET simultaneously captures information from both source and target sentences, enhancing performance. To emulate human evaluation, it adopts diverse training objectives: the estimator directly predicts the quality score of machine-generated text, while the ranking model optimizes by minimizing the distance between high-quality outputs and references, and maximizing the distance for lower-quality outputs. Both components rely on cross-lingual encoders, ensuring robust assessment across languages.

A.4 LLM-based Evaluation

GPTScore: An evaluation framework leveraging generative pre-trained models, such as Flan-T5, to assess text quality. It operates on the premise that high-quality generated text is assigned higher probabilities by the model when conditioned on a given instruction and context. Unlike traditional form-filling approaches, GPTScore frames evaluation as a conditional generation task, capturing the nuanced likelihood of coherent and contextually appropriate outputs.

LLM-as-judges: We harness the evaluative power of state-of-the-art LLMs—including GPT-4.1, LLaMA-3.3, DeepSeek V3, and Gemini 2.5—by designing prompts that enable them to function as reasoning-driven, discourse-aware evaluators. These models provide holistic assessments comparable to expert human reviewers. To maintain consistency and reproducibility, we utilize standardized prompt templates for both summarization and translation tasks.

Prompt Template

System role: You are an expert *{TASK}* evaluator.

Task: Rate the MACHINE *{SUMMARY/TRANSLATION}* strictly on *{ASPECT}*, defined as: *{ASPECT_DEFINITION}*.

Scoring Rubric:
• Use a continuous scale from 0.0 (completely inadequate) to 1.0 (perfect).
• Any decimal allowed (e.g., 0.23, 0.46, 0.71)

Instructions:
• Use (ARTICLE / HUMAN SUMMARY for TS, SOURCE / HUMAN TRANSLATION for MT) as a reference
• Evaluate only on *{ASPECT}*—ignore other criteria
• Output exactly one numeric value between 0.0 (worst) and 1.0 (best).
• No extra text, rationale, or comments

Input:
ARTICLE / SOURCE
HUMAN SUMMARY / TRANSLATION
MACHINE SUMMARY / TRANSLATION

Output:

B Paraphrasing Prompt

To guarantee transparency and reproducibility, we present the exact prompt employed with GPT-4.1 to produce paraphrased summaries.

Prompt Template

System Role: You are an expert paraphraser specialized in rewriting text without changing its meaning.

Task: Paraphrase the given HUMAN SUMMARY so that the core meaning remains unchanged, but the wording and structure are substantially different.

Instructions:
• Preserve all factual and semantic content exactly.
• Use different vocabulary and sentence structures.
• Keep the paraphrase roughly the same length as the original.
• Maintain the same tone and style as the original.
• Do not add, remove, or infer any information.

• Do not explain your reasoning or include any comments.
• Output exactly one paraphrased summary, and nothing else.

Input:
HUMAN SUMMARY: {HUMAN_SUMMARY}

Output:

C Additional Analysis

C.1 System-level Analysis

Table 7 reports Kendall-tau correlations at the system level, reflecting how well automatic metrics align with human judgments in ranking systems. Results reveal remarkable consistency, with most metrics achieving perfect correlations, underscoring their reliability in system ranking. Interestingly, the scores are uniform across all the dimensions assessed, reflecting a consistent concordance between metrics and human evaluations.

Only a handful of metrics—such as SacreBLEU, LEPOR, ROUGE, ChrF, NIST, and BARTScore—show minor deviations across the evaluated dimensions, yet the overall picture confirms the strong effectiveness of automatic metrics in capturing system-level performance.

Metric(s)	TS	MT	Overall
Lexical Overlap			
BLEU, sentBLEU, GLEU, METEOR, ChrF++	1.00	1.00	1.00
SacreBLEU, LEPOR, ROUGE, ChrF	0.33	1.00	0.55
NIST	1.00	0.33	0.78
Embedding-based			
BERTScore, MoverScore, Emb. Avg, Greedy, Vec. Ext., LaBSE, LASER	1.00	1.00	1.00
BARTScore	0.33	1.00	0.55
Neural Learned			
BLEURT, COMET	1.00	1.00	1.00
LLM-based Evaluators			
GPTScore	-0.33	-0.33	-0.33
Llama-3.3, DeepSeek-V3, GPT-4.1 (Eng/Native), Gemini 2.5 Flash	1.00	1.00	1.00

Table 7: System-level Kendall correlations with human judgments.

C.2 Compression Effect

Figure 7 represents the relationship between summary compression (ratio of article tokens to human summary tokens) and average human evaluation scores. Individual points represent observations, the red line shows the linear regression, and the shaded region indicates the 95% confidence interval. The positive slope (0.007) imply a modest preference for summaries that are relatively more concise than their source texts.

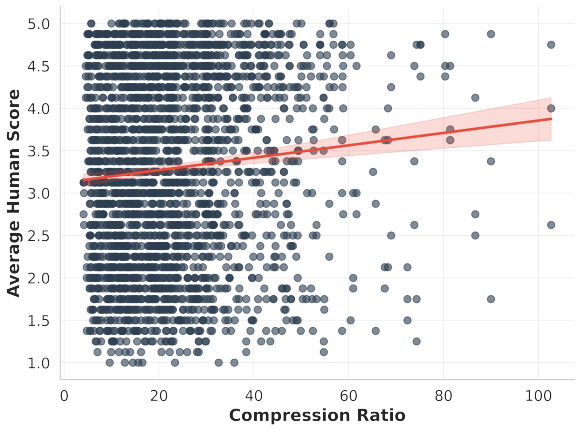


Figure 7: Impact of summary compression on human evaluation scores.

C.3 Entity Noise Variants

Figure 8 shows the relative changes in Pearson correlation between human and COMET scores under different types of entity noise. In MT, entity perturbations substantially reduce correlations, with entity removal causing the steepest declines in both *Adequacy* and *Fluency*. In contrast, for TS, entity noise has a negligible impact, even yielding slight improvements in correlation.

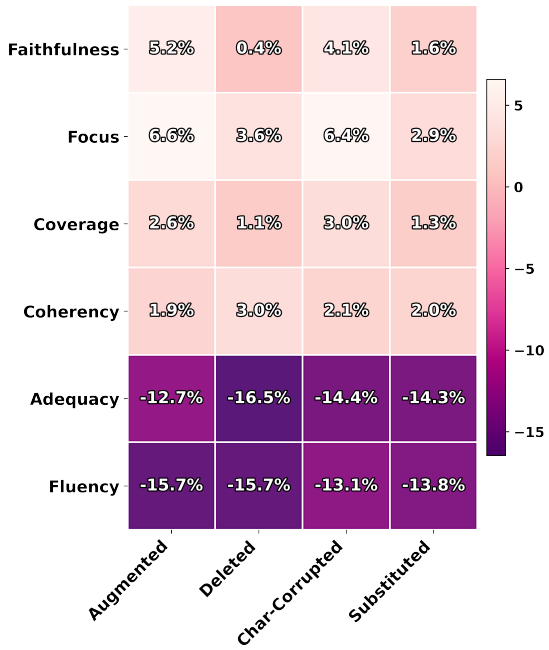


Figure 8: Relative shifts in human–COMET correlations across distinct entity noise conditions.

C.4 Additional Metric Sensitivity

Figure 9 presents the relative ROUGE score drops across languages under various noise perturbations. In MT, shuffling induces the most severe degradation, whereas in TS, inserting off-topic sentences

causes the largest decline. Similar to COMET, the effect of negation varies by language: Tamil and Telugu show marked sensitivity, while other languages remain comparatively robust. As anticipated, ROUGE is overall more susceptible to noise than COMET.

Figure 10 illustrates the relative decreases in human-ROUGE correlations across different evaluation aspects. For MT, entity perturbations most strongly impact *Adequacy*, whereas synonym replacements chiefly reduce *Fluency*. In TS, consistent with COMET's trend, truncating key sentences notably lowers correlations for *Faithfulness*, *Focus*, and *Coverage*, while *Coherency* unexpectedly improves. Overall, *Focus* emerges as the most robust aspect, and *Adequacy* as the least stable.

Figure 11 shows the relative BERTScore declines across languages under diverse perturbations. In MT, shuffling remains the most detrimental, while in TS, character-level edits cause the largest drop. Mirroring COMET's behavior, negation effects are language-dependent: Tamil and Telugu again demonstrate heightened vulnerability. As expected, ROUGE is more noise-sensitive than BERTScore.

Figure 12 highlights the decreases in human–BERTScore correlations by evaluation aspect. In MT, shuffling yields the sharpest reductions in both *Adequacy* and *Fluency*, whereas in TS, negation uniformly lowers all correlations. Collectively, paralleling ROUGE, *Focus* proves most resilient, while *Adequacy* remains the weakest dimension.

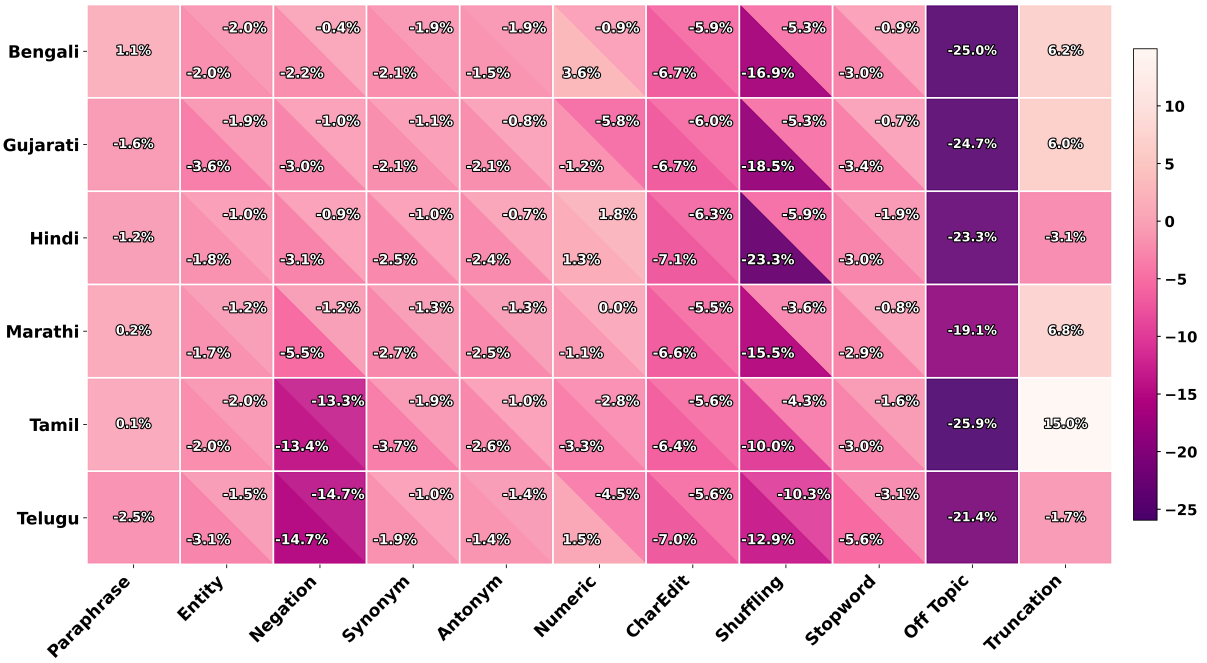


Figure 9: Relative ROUGE score drops (%) across languages under diverse perturbations (top-right: TS; bottom-left: MT).

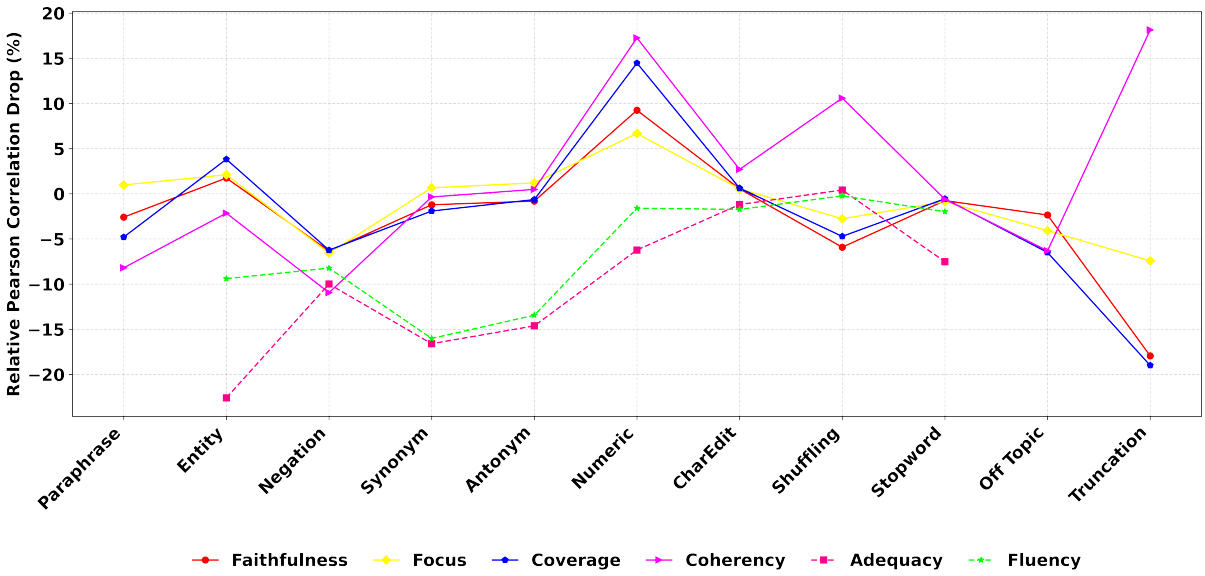


Figure 10: Relative Pearson correlation drops (%) across evaluation aspects under perturbations

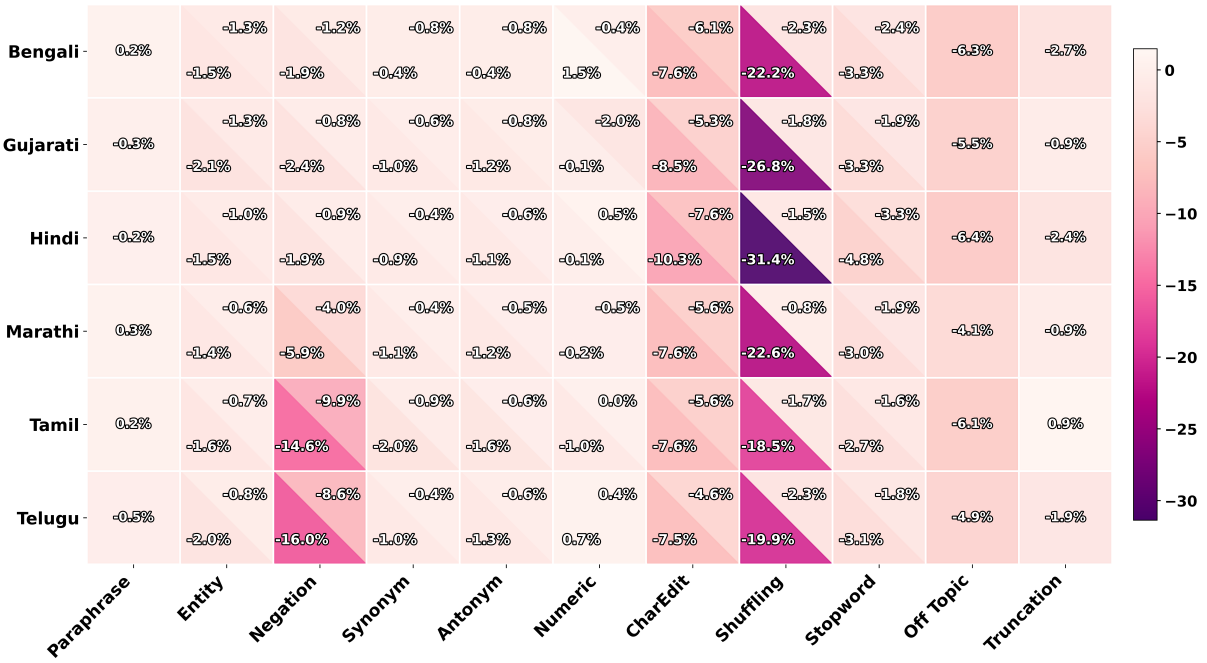


Figure 11: Relative BERTScore score drops (%) across languages under diverse perturbations (top-right: TS; bottom-left: MT).

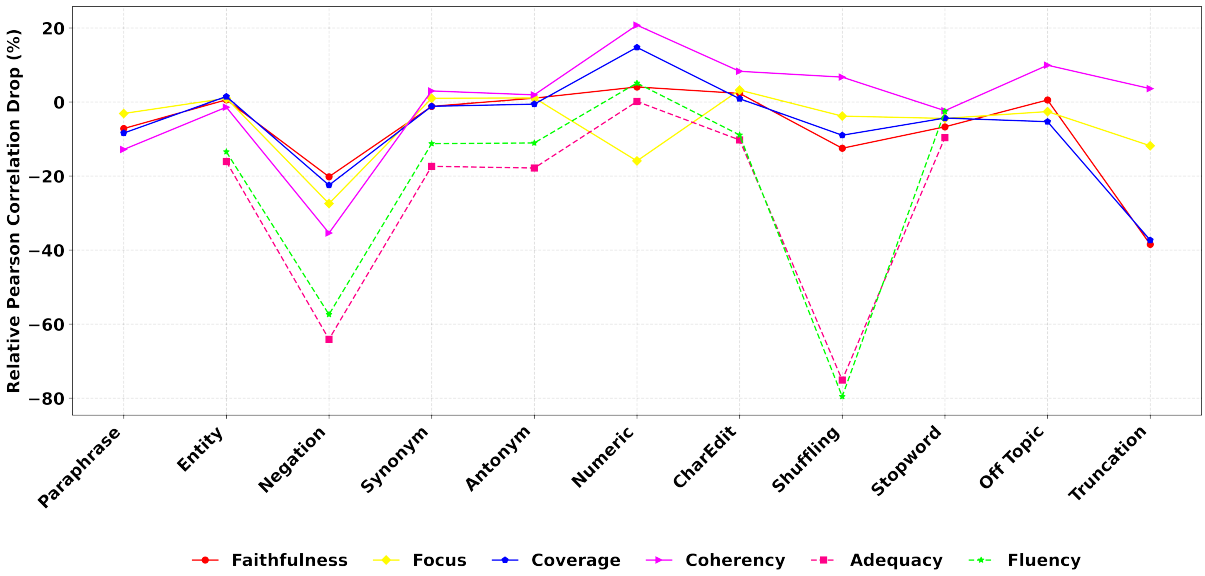


Figure 12: Relative Pearson correlation drops (%) across evaluation aspects under perturbations

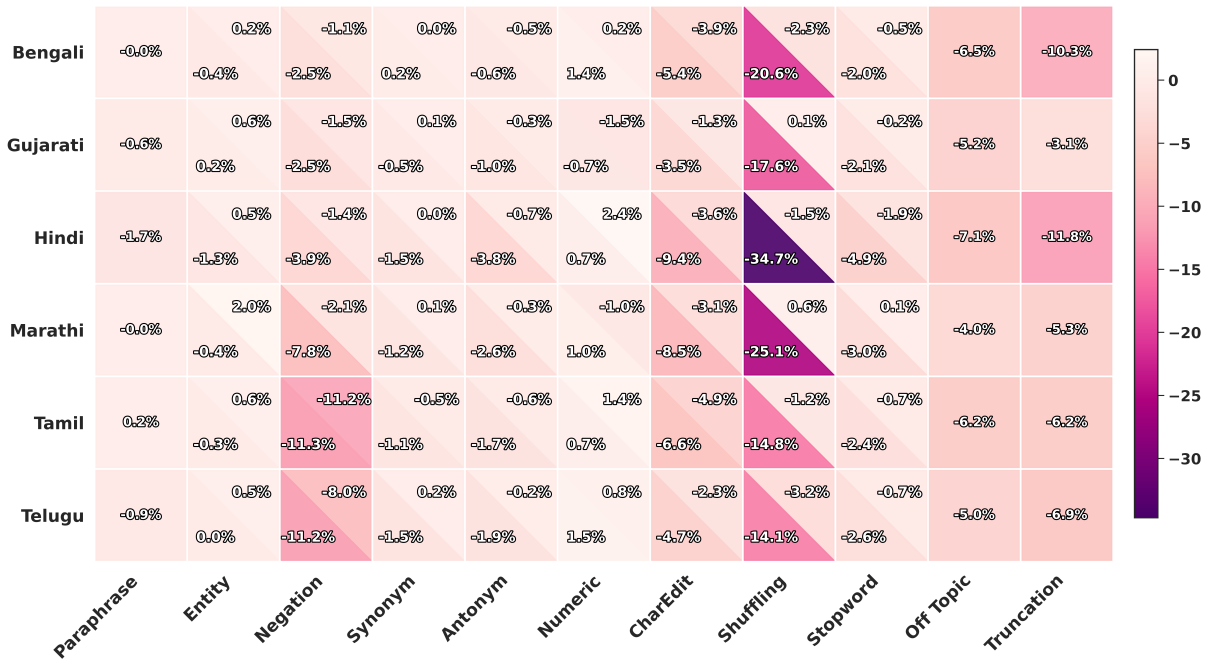


Figure 13: Relative COMET score drops (%) across languages under diverse perturbations (top-right: TS; bottom-left: MT).

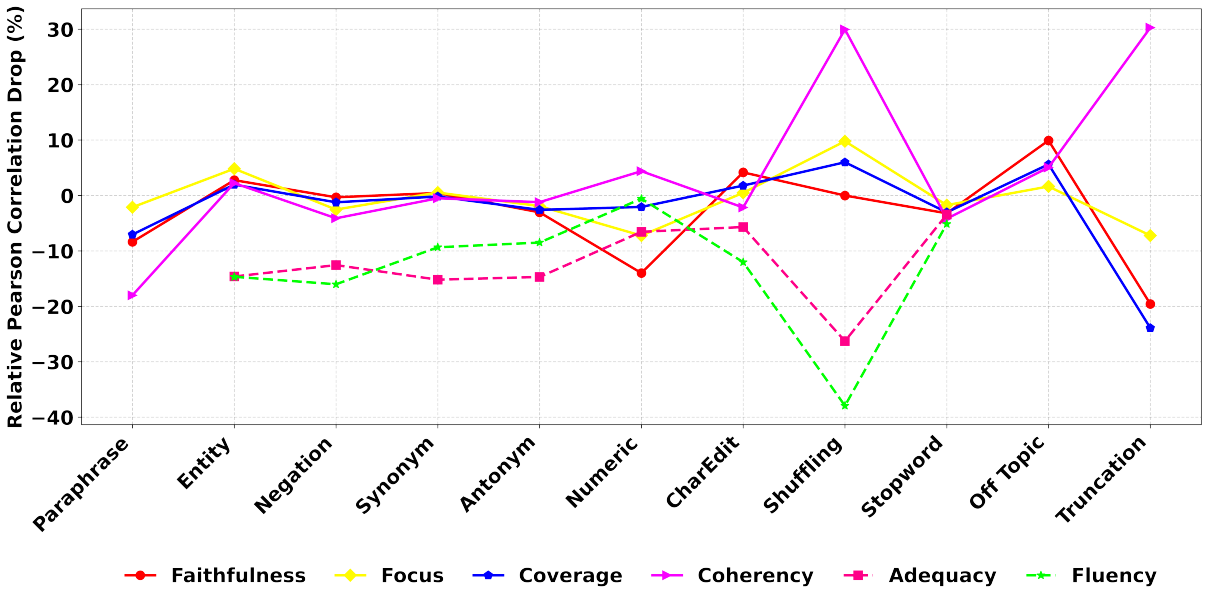


Figure 14: Relative Pearson correlation drops (%) across evaluation aspects under perturbations