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

Amir Hossein Yari, Kalmit Kulkarni, Ahmad Raza Khan, Fajri Koto

I. MOTIVATION & ITEM BENCHMARK

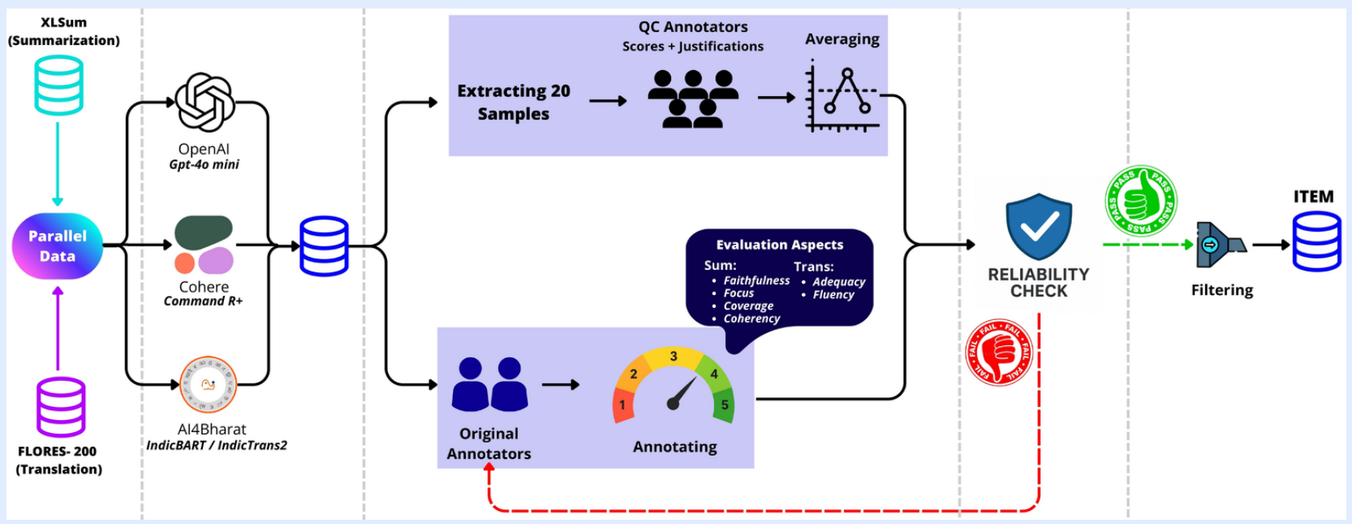
- Automatic evaluation metrics for MT and summarization are largely validated on high-resource languages, leaving their reliability for linguistically diverse Indian languages insufficiently understood.
- Existing evaluations often overlook fine-grained quality aspects and robustness under perturbations, limiting trustworthy assessment of generation quality.
- Introduce ITEM, a benchmark for fine-grained evaluation of MT and summarization metrics in six major Indian languages.

Quality Control:

- Structured QC protocol using 20 representative samples evaluated by 5 independent expert annotators per language and task, with both numerical scores and written justifications.
- Annotator quality measured as the proportion of ratings within ±1 point of expert reference scores across all evaluation aspects.
- Inter-annotator agreement analysis ensured consistency and reliability of human judgments.

Dataset Statistics:

- 2,571 source–summary pairs
- 2,604 source–translation pairs



End-to-End process of dataset creation

Language	Worker 1		Worker 2	
	Sum.	Trans.	Sum.	Trans.
Hindi	81.3	92.5	86.3	85.0
Bengali	80.0	77.5	77.5	85.0
Tamil	85.0	82.5	76.3	82.5
Telugu	92.5	87.5	87.5	87.5
Marathi	76.3	77.5	82.5	87.5
Gujarati	87.5	87.5	91.3	90.0

Annotator quality (%)

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

Inter-annotator Pearson correlations

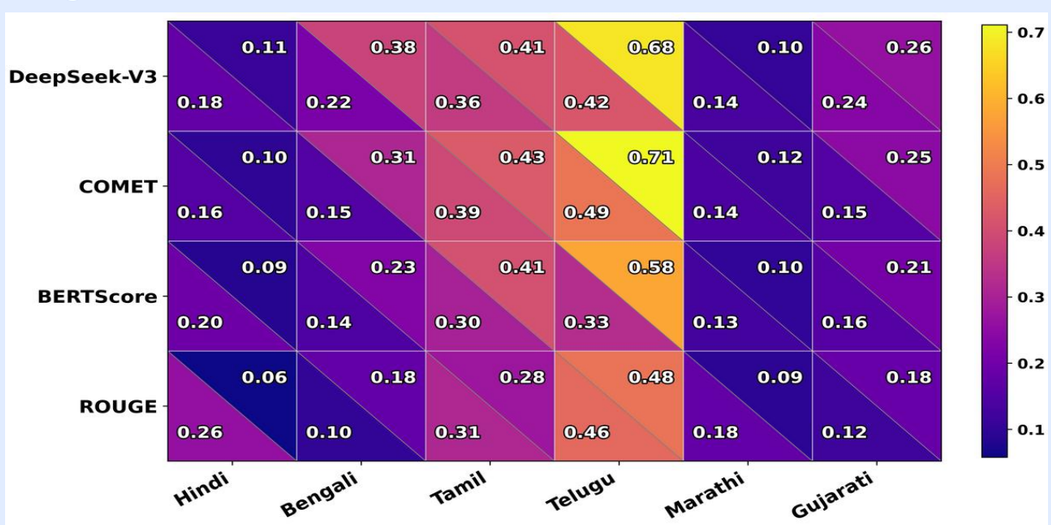
II. MAIN RESULTS & FINDINGS

- Robust outlier detection using Median Absolute Deviation (MAD), Samples with robust z-score (> 3.5) were treated as outliers.
- **Segment-level Analysis:** LLM-based evaluators, especially DeepSeek-V3, show the strongest human alignment; TS metrics better capture content dimensions, while MT metrics better reflect fluency.

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
xCOMET	-0.032 / -0.026	-0.006 / -0.006	-0.025 / -0.02	-0.006 / -0.006	0.286 / 0.281	0.334 / 0.338	0.092 / 0.094
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
MetricX-24 (reference-free)	-/-	-/-	-/-	-/-	**0.311 / 0.298**	0.375 / 0.361	0.343 / 0.33
GEMBA (reference-free)	-/-	-/-	-/-	-/-	0.295 / 0.263	0.366/ 0.354	0.331 / 0.309

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

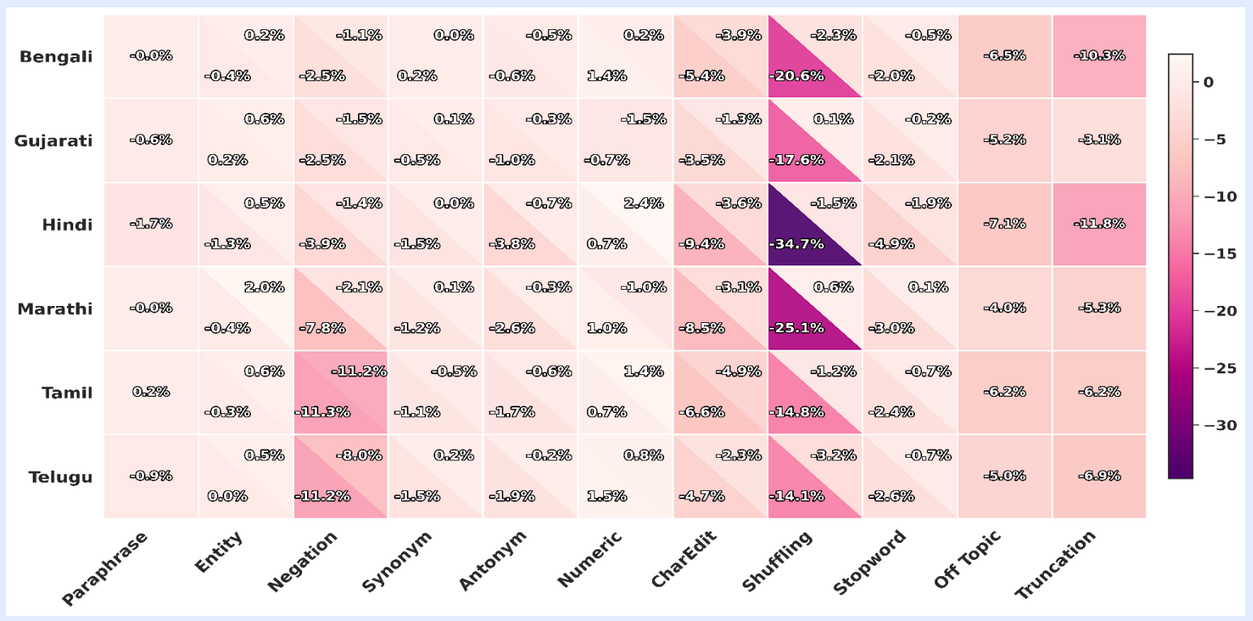
- **Language Effects:** Metric reliability varies by language; Telugu shows the strongest correlations, while Marathi in TS and Hindi in MT are least reliable.



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

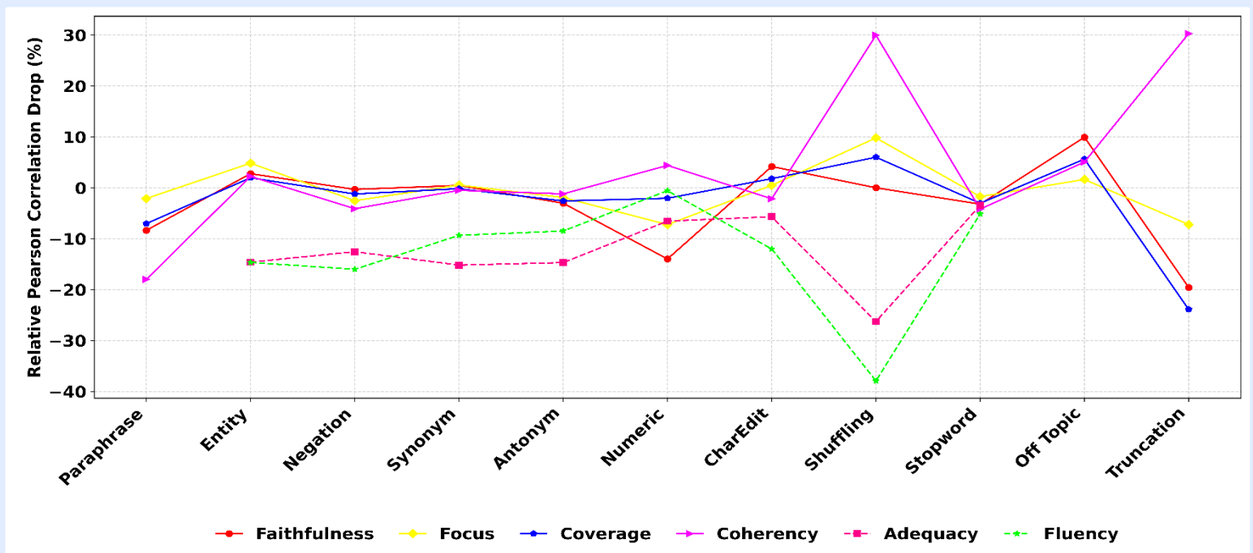
III. ROBUSTNESS ANALYSIS

- GPT-4.1 generated paraphrased summaries, validated by two native speakers for meaning preservation, grammaticality, clarity, and detail consistency.
- Metrics were tested under semantic, lexical, structural, and summarization-specific noise to assess reliability beyond surface similarity.
- COMET is robust to paraphrasing and synonym replacement, but sensitive to structural and meaning-altering noise.
- **Language & Aspect Sensitivity:** Hindi is most vulnerable and Gujarati most resilient; Focus is most stable, while Fluency is least robust.



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

- **Task-specific Effects:** Hindi is most vulnerable and Gujarati most resilient; Focus is most stable, while Fluency is least robust.



Relative Pearson correlation drop (%) between human judgments and COMET across evaluation aspects under perturbations.

CHECK OUT THE PAPER !