LRMGS: A Language-Robust Metric for Evaluating Question Answering in Very Low-Resource Indic Languages

Anuj Kumar[1], **Satyadev Ahlawat**[2], **Yamuna Prasad**[1], **Virendra Singh**[3]
[1]Department of Computer Science & Engineering, Indian Institute of Technology Jammu, India
[2]Department of Electrical Engineering, Indian Institute of Technology Jammu, India
[3]Department of Electrical Engineering, Indian Institute of Technology Bombay, India
{anuj,satyadev.ahlawat,yamuna.prasad}@iitjammu.ac.in, viren@ee.iitb.ac.in

Abstract

Reliable evaluation of Question Answering (QA) systems in low-resource Indic languages poses a significant challenge due to the limited availability of annotated datasets, linguistic diversity, and the lack of suitable evaluation metrics. Languages such as Sindhi, Manipuri, Dogri, Konkani, and Maithili are particularly underrepresented, creating difficulty in assessing Large Language Models (LLMs) on QA tasks. Existing metrics, including BLEU, ROUGE-L, and BERTScore, are effective in machine translation and high-resource settings; however, they often fail in low-resource QA due to score compression, zero-inflation, and poor scale alignment. To overcome this, the Language-Robust Metric for Generative QA (LRMGS) is introduced to capture semantic and lexical agreement while preserving the score scale across languages. LRMGS is evaluated across 8 Indic languages and multiple LLMs, consistently demonstrating higher concordance with reference-based chrF++ scores, as measured using the Concordance Correlation Coefficient (CCC). Experimental results indicate that LRMGS provides more accurate discrimination of system performance in languages with very low resources compared to existing metrics. This work establishes a robust and interpretable framework for evaluating QA systems in low-resource Indic languages, supporting more reliable multilingual model assessment.

1 Introduction

India's linguistic landscape is among the richest globally, yet many languages with millions of speakers remain underrepresented in Natural Language Processing (NLP) and continue to be classified as low-resource due to the scarcity of annotated corpora and benchmarks. Large Language Models (LLMs) hold significant promise for addressing this gap by transferring knowledge from high-resource to low-resource languages through cross-lingual pretraining and generation. Models such as GPT-4 (OpenAI et al., 2024) have demonstrated strong performance in tasks including summarization (Pu et al., 2023; Goyal et al., 2023) and question answering (Zhao et al., 2023), although their training and evaluation processes remain predominantly English-centric. As a result, LLMs frequently struggle to generalize effectively across languages (Lai et al., 2023; Zhang et al., 2023; Ahuja et al., 2023), exhibiting substantial performance disparities between proprietary and open-source models (Ahuja et al., 2024). While multilingual pretraining extends generative capabilities to a wider range of languages (Jiang et al., 2024), evaluation efforts remain constrained by benchmarks dominated by understanding-focused tasks with limited generative coverage (Lai et al., 2023; Asai et al., 2023) and by the continued reliance on expensive reference-based annotations. LLM-based evaluation approaches (Liu et al., 2023) provide an emerging alternative; however, these methods often introduce biases such as a preference for longer outputs or self-generated responses (Zheng et al., 2023; Shen et al., 2023).

Although several Indic QA datasets (Clark et al., 2020; Asai et al., 2021; Singh et al., 2025) have contributed to expanding multilingual evaluation, the core challenge remains the lack of effective evaluation methods for languages with very low resources. Prior efforts often relied on translation-based evaluation (Singh et al., 2024; Chollampatt et al., 2025), which is inadequate for QA since the task requires not only fluent generation, factual correctness, grounding in context, and the preservation of key entities and information. Existing reference-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020), and chrF++ (Popović, 2017) fall short in this setting: they compress score ranges, exhibit weak alignment with human judgments, and often produce unstable rankings across systems. More critically, these metrics operate primarily at the sur-

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 66–77

face or semantic similarity level, thereby overlooking factual aspects of QA such as numeric accuracy, consistency of named entities, and hallucinations. As a result, models that generate fluent yet factually incorrect answers may still receive inflated scores. Cross-lingual protocols aim to mitigate certain issues while introducing new risks, including reference leakage, dependency on expensive annotations, and uncertainty regarding the reliability of scorer LLMs for non-English text.

Example from our result
Language: Dogri **System:** GPT-4.1 **Domain:** Politics
Question: गांधी — इरविन पैक्ट भारत दे निम्नलिखित आंदोलनें च कुस आंदोलन कन्नै जुड़े दा हा ?
Translation: Which of the following movements in India was the Gandhi–Irwin Pact associated with?
Reference: गांधी-इरविन पैक्ट दा सरबंध नागरिक अवज्ञा आन्दोलन कन्नै हा।
Translation: The Gandhi–Irwin Pact was associated with the Civil Disobedience Movement.
Output: गांधी — इरविन पैक्ट सविनय अवज्ञा आंदोलन कन्नै जुड़े दा हा।
Translation: The Gandhi–Irwin Pact is associated with the Civil Disobedience Movement.

Metrics	chrF++	BLEU	BERTScore	**LRMGS**
	0.4852	0.0560	0.9338	**0.9290**

To address the evaluation gap in very low-resource Indic languages, this study builds on the L3Cube-IndicQuest benchmark (Rohera et al., 2024), which includes underrepresented languages such as Sindhi, Manipuri, Dogri, Konkani, and Maithili. The proposed **Language-Robust Metric for Generative QA (LRMGS)** is a composite evaluation framework that integrates semantic similarity through pivoted multilingual BERTScore, nugget-level factual coverage, penalties for numeric mismatches, and evidence-faithfulness checks. Human annotation in these languages remains extremely limited due to the scarcity of bilingual experts, script diversity, and the high cost of large-scale annotation, making direct human correlation infeasible at scale. Consequently, chrF++ is employed as a reproducible *reference metric* for assessing score stability and cross-system concordance. The metric operates purely at the character level and functions as a proxy to examine relative consistency across systems, without modeling semantic or factual correctness. This design allows validation of LRMGS in a principled and language-agnostic manner, even in the absence of human evaluation resources.

2 Evaluation Protocol

2.1 Problem Definition

The task considered in this work is the evaluation of QA outputs across eight low-resource Indic languages. Each evaluation instance is represented as a pair (Q, R), where Q denotes the question posed in one of the target languages and R is its gold reference answer. Given a system prediction $\hat{A}$ produced by a LLM, the objective is to define an evaluation function $\mathcal{E} : (R, \hat{A}) \mapsto s \in [0, 1]$, that assigns a score s reflecting the quality of $\hat{A}$ relative to R.

2.2 Evaluation Metric

To overcome the limitations of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020), and chrF++ (Popović, 2017), which approximate $\mathcal{E}$ via lexical or embedding similarity, the **LRMGS** is introduced. It integrates semantic similarity, question-aware nugget coverage, numeric fidelity, and contextual grounding. Formally,

$$\text{LRMGS} = \prod_{\substack{k \in \{\text{BERT}, \\ \text{KC,NUM,EF}\}}} \text{EN-k}(Ren, \hat{A}_{en}, C_{en})^{\lambda_k}, \tag{1}$$

where $\lambda_{\text{BERT}} = 0.9$, $\lambda_{\text{KC}} = 0.8$, and $\lambda_{\text{NUM}} = \lambda_{\text{EF}} = 1$.

Notation. A QA instance is represented as (Q, R), where Q is the Indic question, R the gold answer, and $\hat{A}$ the system prediction. English translations of Q and R are provided, and $\hat{A}$ is translated via IndicTrans2 (Gala et al., 2023) for consistent evaluation. Let Q_{en}, R_{en}, and $\hat{A}en$ denote the English forms of the question, reference, and system output, with context $Cen = c_1, \ldots, c_m$ representing the English question sentences for grounding. For EN-BERTScore, token embeddings $\mathbf{r}i$ and $\mathbf{a}j$ are obtained using RoBERTa-large (Zhang* et al., 2020). For key-nugget coverage (KC) and evidence faithfulness (EF), Sentence-Transformers (Reimers and Gurevych, 2019) encode nuggets $\mathbf{k}i$ and context sentences $\mathbf{c}$. Nuggets correspond to factual clauses segmented from Ren, with attention weights a_i derived via softmax-normalized similarity to Qen, emphasizing the most relevant clauses. Numeric sets N_R and $N\hat{A}$ contain expressions extracted through regular expressions.

Semantic similarity (EN-BERT).

$$\text{EN-BERT}(R_{en}, \hat{A}_{en}) = \frac{1}{|R_{en}|} \sum_{i=1}^{|R_{en}|} \max_j \left|\cos(\mathbf{r}_i, \mathbf{a}_j)\right|. \quad (2)$$

Question-aware nugget attention (EN-KC).

$$\text{EN-KC}(R_{\text{en}}, Q_{\text{en}}) = \frac{\exp\left(\frac{\cos(\mathbf{k}_i, \mathbf{q})}{\eta}\right)}{\sum_{j=1}^{n} \exp\left(\frac{\cos(\mathbf{k}_j, \mathbf{q})}{\eta}\right)}, \quad (3)$$

where $\mathbf{k}_i$ and $\mathbf{q}$ are embeddings of clause c_i (from R_{en}) and question Q_{en}, n is the number of clauses, and η is the temperature controlling attention sharpness. A smaller η yields peaked attention, whereas a larger value smooths the distribution. Top-k nuggets with the highest attention weights are retained as key concepts for coverage computation.

Numeric fidelity (EN-NUM).

$$\text{EN-NUM}(R_{en}, \hat{A}_{en}) = \frac{|N_R \cap N_{\hat{A}}|}{|N_R \cup N_{\hat{A}}|}. \quad (4)$$

A partial penalty is applied when the reference contains numbers; however, the hypothesis does not, as reflected in the implementation.

Evidence faithfulness (EN-EF).

$$\text{EN-EF}(C_{en}, \hat{A}_{en}) = \max_{c \in C_{en}} \cos(\mathbf{a}, \mathbf{c}), \quad (5)$$

which ensures contextual grounding by requiring the generated answer to align semantically with at least one translated question sentence. The full algorithmic implementation of LRMGS is provided in Appendix D.

3 Evaluation and Dataset

Two kinds of evaluation have been done in this work: (1) Meta-evaluation and (2) LLM Comparison.

Meta-evaluation: The ability of LRMGS to substitute conventional reference-based metrics for multilingual QA evaluation is examined by computing its concordance with chrF++ (Popović, 2017), a metric shown to align well with human judgments in multilingual text generation (Singh et al., 2024). The Concordance Correlation Coefficient (CCC) (ccc, 1989) is employed, as it evaluates both precision (Pearson correlation) and accuracy (closeness to the identity line), thereby capturing true agreement rather than only monotonic consistency.

Language	BLEU	BScore	LRMGS
Assamese	0.406	0.0229	0.627
Dogri	0.376	0.0172	0.538
Hindi	0.541	0.0286	0.646
Konkani	0.356	0.0209	0.597
Maithili	0.430	0.0210	0.563
Manipuri	0.413	0.0153	0.580
Sanskrit	0.276	0.0222	0.601
Sindhi	0.569	0.0221	0.642
Average	0.421	0.0213	**0.599**

Table 1: Comparison of correlation-based agreement with chrF++ across metrics for each low-resource Indic language. The results show that LRMGS consistently achieves higher concordance with chrF++ than BLEU and BERTScore(BScore).

System	BLEU	BScore	LRMGS
Airavata-7B	0.361	0.017	0.539
Aya-23-8B	0.515	0.012	0.586
BLOOMZ-7B	0.736	0.006	0.362
GPT-4.1	0.390	0.024	0.571
Gemma-2-9B-it	0.498	0.020	0.565
Llama-3.1-8B	0.445	0.018	0.542
Mistral-7B	0.212	0.005	0.430
OpenHathi7B-Hi	0.101	0.006	0.486
Qwen2.5-7B-Inst.	0.375	0.013	0.502
Yi-1.5-9B-Chat	0.648	0.001	0.311
Average	0.418	0.013	**0.500**

Table 2: System-level comparison of correlation-based agreement with chrF++ across evaluation metrics. LRMGS achieves the highest average concordance (0.500), outperforming BLEU and BERTScore(BScore) across diverse multilingual systems.

Formally, for score sets $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$,

$$\rho_c = \frac{2\rho\,\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}, \quad (6)$$

where ρ denotes the Pearson correlation coefficient between X and Y, while $\mu_\cdot$ and $\sigma_\cdot^2$ denote the mean and variance, respectively. A value of $\rho_c = 1$ indicates perfect concordance.

CCC is reported at two levels of granularity, consistent with the evaluation protocol: (i) *language-level*, where for each language ℓ, CCC is computed between LRMGS and chrF++ over all samples belonging to ℓ; (ii) *system-level*, where for each system s, CCC is computed between LRMGS and chrF++ over all samples generated by s without

	Assamese	Dogri	Hindi	Konkani	Maithili	Manipuri	Sanskrit	Sindhi
OpenHathi-7B-Hi-Base	0.051	0.199	0.248	0.153	0.209	0.041	0.159	0.051
Yi-1.5-9B-Chat	0.056	0.052	0.045	0.044	0.049	0.046	0.044	0.051
BLOOMZ-7B1-mt	0.118	0.129	0.162	0.088	0.124	0.04	0.121	0.067
Aya-23-8B	0.17	0.224	0.238	0.222	0.239	0.044	0.22	0.149
Mistral-7B	0.218	0.224	0.252	0.209	0.244	0.085	0.166	0.181
Airavata-7B	0.253	0.258	0.268	0.229	0.27	0.037	0.242	0.2
Qwen2.5-7B-Instruct	0.28	0.294	0.305	0.289	0.305	0.136	0.278	0.269
Gemma-2-9B-it	0.29	0.284	0.331	0.278	0.309	0.169	0.266	0.254
Llama-3.1-8B-Instruct	0.282	0.327	0.348	0.314	0.336	0.215	0.297	0.279
GPT-4.1	**0.411**	**0.394**	**0.434**	**0.39**	**0.422**	**0.267**	**0.361**	**0.38**

Table 3: System × language matrix of LRMGS scores with prompts in English. The results highlight consistent cross-lingual trends, with GPT-4.1 achieving the highest scores across all eight Indic languages.

pre-averaging.

LLM Comparison: Using LRMGS, multilingual QA evaluation across ten LLMs shows stronger agreement in medium-resource languages like Hindi and Assamese, while very low-resource ones such as Sindhi and Dogri reveal large performance gaps. GPT-4.1 achieves the best overall results, though open-source models like Gemma-2-9B-IT and LLaMA-3.1-8B-Instruct are competitive and surpass larger proprietary systems in some cases. These results underline the strengths of proprietary models while highlighting the growing potential of open-source alternatives. Further details about LLMs and prompt are given in Appendix E.

Dataset: The study uses the L3Cube-IndicQuest dataset (Singh et al., 2025), containing 4,000 QA pairs across 20 languages, each with 200 questions from five domains. The questions were originally authored in English, manually verified for correctness, and subsequently translated into the other Indic languages. For this work, eight low-resource languages, Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi, are selected to examine multilingual evaluation under data scarcity. Further details about the dataset are mentioned in Appendix A.

4 Results

Tables 1 and 2 establish the reliability of LRMGS as an evaluation protocol across Indic languages and multilingual systems. Unlike BLEU and BERTScore, which underperform in very low-resource scenarios, LRMGS demonstrates stronger concordance with chrF++ both across languages and across systems. This robustness is most apparent for Dogri and Manipuri, where surface-based metrics fail to capture semantic fidelity or factual adequacy. At the system level, BLEU occasionally aligns with chrF++ for certain models yet fluctuates sharply for others, while BERTScore remains uniformly weak. LRMGS provides stable and interpretable agreement, confirming its suitability for benchmarking multilingual QA tasks in low-resource conditions.

To illustrate how individual components of LRMGS contribute to the final score, Table 4 presents representative GPT-4.1 outputs. These examples illustrate how the metric captures semantic fidelity, factual grounding, numeric consistency, and alignment with contextual evidence, in contrast to conventional metrics that rely solely on surface overlap. High LRMGS values correspond to fluent and factually correct paraphrases, whereas mid-range or low scores indicate partial factual omission or semantic drift. This qualitative behavior explains the improved discriminative reliability of LRMGS across languages and systems.

The results reveal several key insights. First, LRMGS exhibits greater score stability across Indic languages of varying resource levels, maintaining consistent scale alignment even in the presence of translation noise. Second, correlation with chrF++ confirms that LRMGS preserves rank consistency across systems while extending evaluation to factual and contextual dimensions that chrF++ does not capture. Third, LRMGS demonstrates stronger discriminative power in identifying fine-grained differences among LLMs, particularly between proprietary and open-source models. These insights validate the interpretability and robustness of LRMGS as a framework for evaluation.

Although the metric employs English pivoting through translation, the direct application of En-

Table 4: Illustrative GPT-4.1 examples highlighting the contribution of each LRMGS component (EN-BERT, EN-KC, EN-NUM, EN-EF) in evaluating semantic, factual, numeric, and contextual faithfulness.

Language: Sanskrit **Question (EN):** In which caves is the Kailasha temple located? **Reference (EN):** The Kailasha temple is located in the Ellora caves. **Output (EN):** The Kailash Temple is located in the Ellora Caves. **Scores:** BLEU = 0.427 chrF++ = 0.803 EN-BERT = 0.970 EN-KC = 0.928 EN-NUM = 1.000 EN-EF = 0.978 LRMGS = 0.896 **Interpretation:** A fluent paraphrase preserving all factual elements. Despite moderate BLEU, both **EN-BERT (Eq. 2)** and **EN-EF (Eq. 5)** remain near 1.0, reflecting semantic and contextual fidelity. LRMGS correctly assigns a high score (0.896) while surface metrics undervalue it.
Language: Assamese **Question (EN):** How did Hamlet's father die? **Reference (EN):** Hamlet's father was killed by his brother Claudius with a drink laced with poison. **Output (EN):** Hamlet's father was killed by his brother Claudius. **Scores:** BLEU = 0.008 chrF++ = 0.187 EN-BERT = 0.887 EN-KC = 0.513 EN-NUM = 1.000 EN-EF = 0.942 LRMGS = 0.495 **Interpretation:** A semantically correct yet contextually reduced answer. High **EN-BERT** yet lower **EN-KC** indicate factual alignment with partial omission of narrative context. This illustrates how **Equation (3)** penalizes incomplete nugget coverage, leading to a mid-range LRMGS score.

glish metrics, such as BLEU or ROUGE, to translated text is unreliable. Translation artifacts frequently alter lexical structure and token boundaries, reducing the validity of token-based similarity. LRMGS mitigates these effects by performing semantic alignment on the pivoted text and incorporating numeric and evidence-based checks that remain stable under translation noise. This hybrid design enables reliable cross-lingual evaluation while preserving the linguistic and factual characteristics of the original Indic content.

Having validated the metric, Table 3 applies LRMGS to compare large language models across eight Indic languages. The results reveal clear variation across models and languages. Newer instruction-tuned architectures, including LLaMA-3.1-8B-Instruct and Gemma-2-9B-it, consistently outperform earlier baselines such as BLOOMZ and Yi-1.5-9B-Chat, indicating the benefit of recent training improvements for low-resource QA. Airavata-7B and Qwen2.5-7B-Instruct achieve competitive scores, although challenges persist for Manipuri and Sindhi, where most models display sharp performance degradation. GPT-4.1 attains the highest LRMGS values across all languages, reflecting both the disparity between proprietary and open-source systems and the capability of LRMGS to capture these nuanced differences. Comprehensive results for BLEU, ROUGE-L, chrF++, and BERTScore, along with visualizations, are provided in Appendix C.

5 Conclusion

This work introduced LRMGS, a composite evaluation metric designed for generative QA in very low-resource Indic languages. By combining semantic similarity, nugget-level coverage, numeric consistency, and evidence faithfulness, LRMGS captures both factual and contextual dimensions of QA quality. Experiments across eight Indic languages show that LRMGS consistently achieves higher concordance with chrF++ compared to BLEU and BERTScore. The results highlight its robustness in ranking multilingual systems and its ability to reveal performance gaps in underrepresented languages. LRMGS thus provides a reliable and interpretable framework for benchmarking QA systems in low-resource settings.

Limitations

This work has several limitations. First, LRMGS relies on translation to English through IndicTrans2, which may introduce translation errors and slightly influence the resulting evaluation scores. Second, the evaluation is limited to eight Indic languages due to the availability of suitable datasets, leaving a substantial number of low-resource languages unexamined. Third, meta-evaluation is performed against chrF++ rather than direct human judgments for all languages, thereby constraining the strength of conclusions regarding alignment with human evaluations.

References

1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. 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 2598–2637.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *Preprint*, arXiv:2305.14857.

Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. Cross-lingual evaluation of multilingual text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul 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. *Transactions on Machine Learning Research*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

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

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992.

Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context. *arXiv preprint arXiv:2409.08706*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. INDIC QA BENCHMARK: A multilingual benchmark to evaluate question answering capability of LLMs for Indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

A Dataset

Dataset size and length characteristics. Table 5 summarizes per–language dataset statistics. Each language has 200 QA pairs, and average *question* length ranges from 7.56 (Sanskrit) to 11.22 tokens (Dogri), while average *answer* length ranges from 19.32 (Sanskrit) to 30.50 tokens (Dogri/Sindhi). Across languages, answers are roughly 2.5×–3× longer than questions, indicating that systems must handle short prompts with substantially longer generations.

Language	Samples	Avg Q tokens	Avg A tokens
Assamese	200	8.55	23.08
Dogri	200	11.22	30.5
Hindi	200	11.01	29.8
Konkani	200	8.3	22.16
Maithili	200	11.08	30.48
Manipuri	200	8.62	23.36
Sanskrit	200	7.56	19.32
Sindhi	200	11.02	30.48

Table 5: Dataset statistics per language: number of samples and average token lengths of questions/answers.

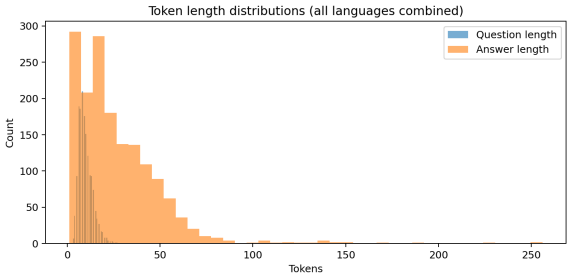

Figure 1: Token–length distributions aggregated across all languages. Questions are short and tightly clustered; answers are longer and right–skewed with a long tail.

Figure 1 shows the combined token–length distributions over all languages. Question lengths are tightly concentrated in the 5–15 token range, whereas answer lengths exhibit a broader, right–skewed distribution with a long tail (occasionally exceeding 200 tokens). The separation between the two histograms suggests limited confounding between prompt length and response length, and the heavy–tailed answers motivate clause–level scoring and attention mechanisms.

B Experimental Setup and Metrics

Experiments are conducted across eight low-resource Indic languages: Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi, using 200 QA pairs per language. For each (question, reference) pair, model outputs are generated and evaluated using both automatic reference-based metrics and human-aligned LLM ratings. All experiments are inference-only, with no model updates or gradient computations. Results are reported per language and per system, followed by correlation analysis using Pearson, Spearman, and Kendall's τ.

Generation Settings. Inference is performed using the `transformers` library in `float16` precision with `device_map=auto` across **2× NVIDIA V100 PCIe 32 GB** GPUs. Decoding uses deter-

ministic **greedy search** (do_sample=false) with a limit of 128 new tokens. Tokenizers use left padding and truncation, defaulting to the eos_token when the pad_token is undefined. Random seeds are fixed to 42 for Python, NumPy, and PyTorch to ensure reproducibility. Batch size is one, and no gradients are computed.

Automatic Metrics. Evaluation includes **BLEU**, **ROUGE-L** (LCS F1), **chrF++** (β=2), and English-projected **BERTScore (F1)**, along with the proposed **LRMGS** metric that captures semantic and factual grounding in multilingual QA. BLEU scores are computed using a maximum of four-gram overlap (**BLEU-4**) with standard smoothing (method 1). Lower-order BLEU variants (1–3) were additionally examined for consistency, and system-level rankings remained stable across all configurations. All BLEU results reported in the tables correspond to BLEU-4.

C Visualization plots and Example Analysis

Analysis of Metric Correlation with chrF++. Figures 2–8 provide a detailed comparison of how different metrics correlate with chrF++ across languages, systems, and individual sentences.

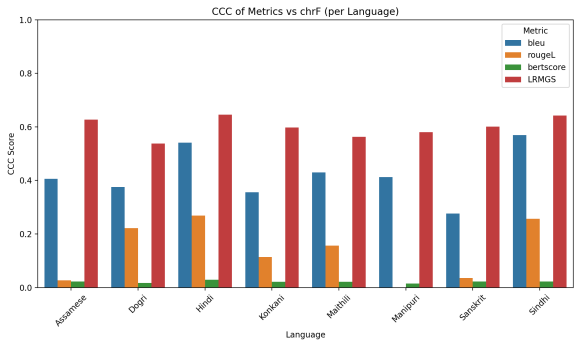

Figure 2: CCC of metrics vs. chrF++ across Indic languages. LRMGS consistently achieves the highest correlation.

Language-level correlation. Figures 2 and 3 report the concordance correlation coefficient (CCC) *between chrF++ and each other metric* (BLEU, ROUGE-L, BERTScore, LRMGS) across eight Indic languages. BLEU exhibits *moderate* agreement with chrF++ (≈0.28–0.57): e.g., Assamese ≈0.41, Dogri ≈0.38, Hindi ≈0.54, Konkani ≈0.36, Maithili ≈0.43, Manipuri ≈0.41, Sanskrit ≈0.28, Sindhi ≈0.57. ROUGE-L remains *weak* (≈0.00–0.27) and even slightly negative in Manipuri, reflecting brittle span matching under rich morphology and orthographic variation. BERTScore is *near zero* everywhere (≈0.015–0.03), indicating that sentence-level embedding similarity poorly tracks character overlap in these low-resource settings (saturation and insensitivity to small lexical differences). In contrast, LRMGS is *uniformly higher and tightly clustered* (≈0.54–0.65)—e.g., ≈0.63 for Assamese, ≈0.65 for Hindi, ≈0.60 for Konkani/Sanskrit, ≈0.64 for Sindhi—demonstrating robust concordance with chrF++ across scripts and families. The bar plot reiterates this: LRMGS dominates BLEU/ROUGE-L/BERTScore for every language, with especially strong margins in Assamese, Hindi, and Sindhi.

Why these patterns arise. BLEU's token-level n-gram matching favors languages with relatively stable tokenization (e.g., Hindi, Sindhi), while its effectiveness declines for morphologically rich or compound-heavy languages such as Sanskrit and Konkani, where surface forms diverge substantially from reference expressions. ROUGE-L's reliance on the longest common subsequence makes it highly sensitive to word order and segmentation, both of which vary considerably across Indic scripts, thereby reducing correlation with human judgments (CCC). BERTScore frequently saturates at high cosine similarity values, leading to limited variance and consequently weaker alignment with chrF++. In contrast, LRMGS integrates semantic similarity, question-aware nugget coverage, numeric fidelity, and contextual grounding, generating scores that vary meaningfully with factual and semantic alignment, thereby exhibiting stronger concordance with chrF++ patterns.

System-level correlation. Figure 4 presents CCC *between chrF++ and each metric* by model. BLEU is *variable across systems* (higher for some instruction-tuned or stronger decoders, lower for others such as BLOOMZ and Mistral-7B). ROUGE-L remains *uniformly small* (roughly 0.05–0.20), while BERTScore is *near zero* for all systems. LRMGS is *consistently mid-to-high* (typically ≈0.43–0.59) with a narrow spread across model families (Gemma, LLaMA, Qwen, GPT, etc.), indicating stable agreement with chrF++ irrespective of architecture or size.

Takeaways. (i) LRMGS demonstrates the *highest and most consistent* CCC with chrF++ across both languages and systems; (ii) BLEU remains *functional yet inconsistent*, with performance vary-

ing by language morphology and model type; (iii) ROUGE-L and BERTScore serve as *unreliable correlates* of chrF++ under multi-script, low-resource conditions due to segmentation sensitivity in ROUGE-L and embedding saturation in BERTScore.

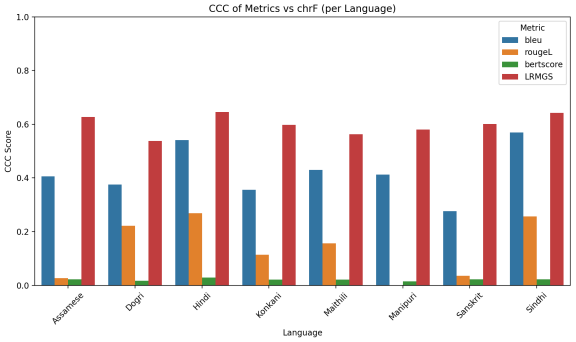

Figure 3: Language-level CCC bar plots comparing metrics with chrF++. LRMGS shows consistent improvements.

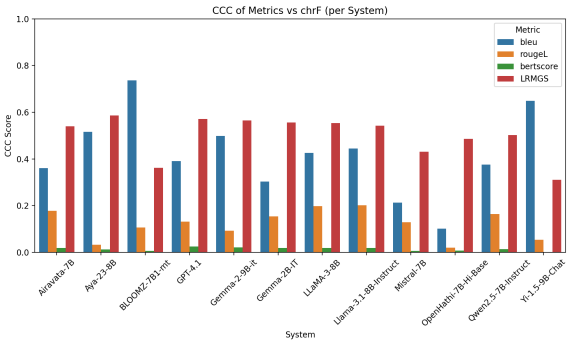

Figure 4: System-level CCC across multiple LLMs. LRMGS remains stable compared to BLEU, ROUGE-L, and BERTScore.

Sentence-level scatter plots. Figures 5–8 examine sentence-level correlations. Figure 5 shows LRMGS vs. chrF++ with a dense positive trend and strong linearity, validating its reliability at fine granularity. Figure 6 shows BLEU vs. chrF++ with weaker and noisier alignment; BLEU often fails to capture quality when chrF is moderate-to-low. Figure 7 shows BERTScore vs. chrF++, where values saturate near 1.0, leading to compressed scores and poor discrimination. Finally, Figure 8 compares chrF++ and LRMGS, highlighting that LRMGS captures semantic fidelity while maintaining correlation with character-level overlaps. This balance explains why LRMGS consistently shows higher concordance across languages and systems.

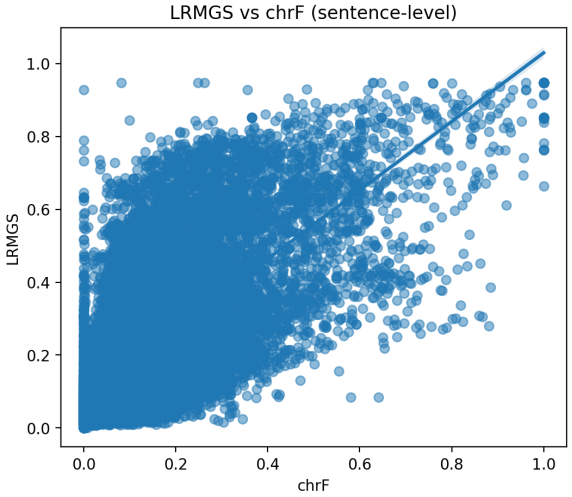

Figure 5: Sentence-level correlation of LRMGS vs. chrF++. Strong positive alignment validates reliability.

Example Analysis. Table 6 presents six GPT-4.1 QA examples across Assamese, Dogri, Maithili, and Manipuri, each showing the gold reference (Indic and English) alongside the model's output (Indic and English) evaluated using four metrics. In the first four rows (two Assamese and two Dogri examples), the model's outputs closely paraphrase the references, maintaining alignment in names, facts, and phrasing. This results in consistently high chrF++ scores (0.875–0.905), moderate BLEU values (0.531–0.809), very high BERT similarities (0.990–0.997, except 0.970/0.981), and strong LRMGS scores (0.896–0.930), collectively indicating strong semantic fidelity and contextual coverage. The Maithili example illustrates a clear failure case: the model hallucinates an unrelated religious ceremony for "FERA," causing chrF++ and BLEU to collapse (0.157/0.012), BERT similarity to drop (0.826), and LRMGS to approach zero (0.018), reflecting both lexical and semantic divergence. The Manipuri example shows partial comprehension yet limited grounding and structural coherence; metrics are mixed (chrF++ 0.304, BLEU 0.268, high BERT 0.949 due to token overlap, and very low LRMGS 0.017), demonstrating that surface similarity can be misleading. LRMGS, in contrast, effectively penalizes unfaithful or non-answering content. Overall, faithful factual matches yield high scores across metrics, whereas semantic errors or off-topic responses are strongly penalized, most notably by LRMGS.

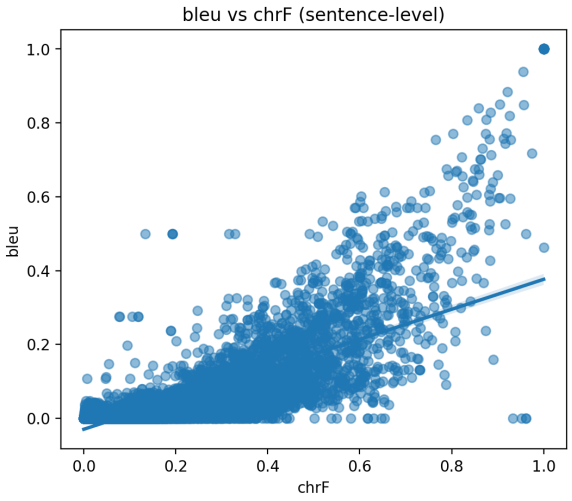

Figure 6: Sentence-level correlation of BLEU vs. chrF++. BLEU shows weaker correlation and noisy behavior.

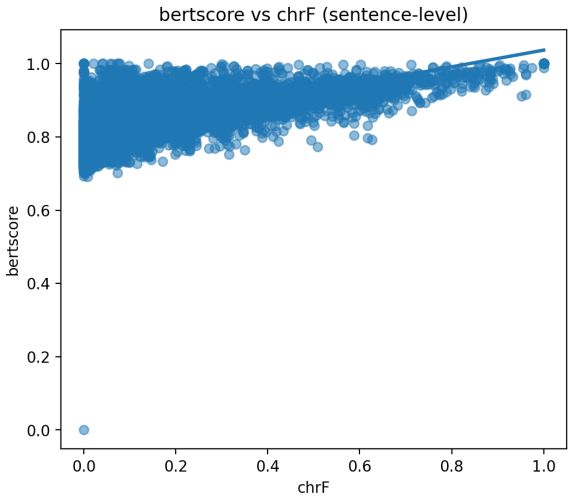

Figure 7: Sentence-level correlation of BERTScore vs. chrF. Scores saturate near 1.0, limiting discrimination.

Algorithm 1: Computation of LRMGS Metric (Symbolic Form)

Input: (Q, R), $\hat{A}$, weights $\{\lambda_{\text{BERT}}, \lambda_{\text{KC}}, \lambda_{\text{NUM}}, \lambda_{\text{EF}}\}$, temperature η

Output: LRMGS $\in [0, 1]$

1. Preprocessing:

Translate using IndicTrans2:

$Q_{en}, R_{en}, \hat{A}_{en} \leftarrow \text{Trans}_{\text{IndicTrans2}}(Q, R, \hat{A})$; split question sentences as $C_{en} = \{c_1, \ldots, c_m\}$.

2. Semantic Similarity (EN-BERT):

$\text{EN-BERT} = \frac{1}{|R_{en}|} \sum_i \max_j |\cos(\mathbf{r}_i, \mathbf{a}_j)|$.

3. Question-Aware Nugget Coverage (EN-KC):

Segment R_{en} into factual clauses $\{c_i\}_{i=1}^n$ and embed $\mathbf{k}_i = \text{ST\_embed}(c_i)$, $\mathbf{q} = \text{ST\_embed}(Q_{en})$, $\hat{\mathbf{a}}_j = \text{ST\_embed}(\hat{A}_{en})$.

Compute nugget attention $a_i = \frac{e^{\cos(\mathbf{k}_i, \mathbf{q})/\eta}}{\sum_j e^{\cos(\mathbf{k}_j, \mathbf{q})/\eta}}$, and compute

$$\text{coverage EN-KC} = \frac{\sum_i a_i \max_j |\cos(\mathbf{k}_i, \hat{\mathbf{a}}_j)|}{\sum_i a_i}.$$

4. Numeric Fidelity (EN-NUM):

$N_R = \text{RegexNums}(R_{en})$, $N_{\hat{A}} = \text{RegexNums}(\hat{A}_{en})$,

$$\text{EN-NUM} = \frac{|N_R \cap N_{\hat{A}}|}{|N_R \cup N_{\hat{A}}|}.$$

5. Evidence Faithfulness (EN-EF):

$\mathbf{a} = \text{ST\_embed}(\hat{A}_{en})$, $\mathbf{c} = \text{ST\_embed}(C_{en})$, $\text{EN-EF} = \max_{c \in C_{en}} \cos(\mathbf{a}, \mathbf{c})$.

6. Aggregation:

$$\text{LRMGS} = (\text{EN-BERT})^{\lambda_{\text{BERT}}} (\text{EN-KC})^{\lambda_{\text{KC}}} (\text{EN-NUM})^{\lambda_{\text{NUM}}} (\text{EN-EF})^{\lambda_{\text{EF}}}.$$

D Evaluation Algorithm (Symbolic)

The overall procedure for computing the Language-Robust Metric for Generative QA (LRMGS) is formalized in Algorithm 1. It integrates four components: semantic similarity (EN-BERT), question-aware keypoint extraction and coverage (EN-KP/EN-KC), numeric consistency (EN-NUM), and evidence faithfulness (EN-EF). The algorithm ensures reproducible evaluation of QA systems under multilingual and low-resource settings.

E Large Language Models and Experimental Setup

For benchmarking, a diverse suite of large language models (LLMs) was employed, encompassing both open-source Indic models and general-purpose multilingual LLMs. All models were evaluated within a unified framework designed to ensure reproducibility and fairness across Indic languages.

Models. The following LLMs were included:

- **Mistral-7B** (causal decoder-only), Hugging Face mistralai/Mistral-7B-v0.1.
- **OpenHathi-7B-Hi-Base**, optimized for Hindi and related Indic languages.
- **Qwen2.5-7B-Instruct**, trained with multilingual instruction-following data.
- **Yi-1.5-9B-Chat**, a decoder-only chat-tuned model.
- **GPT-4.1**, accessed via API, serving as a high-capacity commercial baseline.
- **Gemma-2-9B-it**, Google's instruction-tuned Gemma model with strict chat templates.
- **Airavata-7B**, an Indic-focused model from AI4Bharat using open-instruct style prompting.

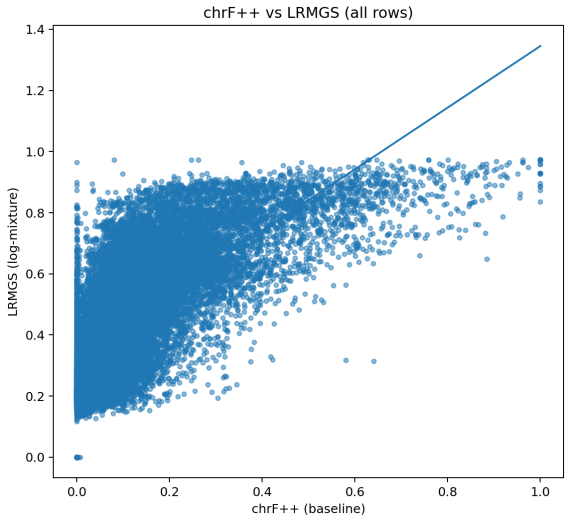

Figure 8: Comparison of chrF++ vs. LRMGS at the sentence level. LRMGS maintains correlation while capturing semantic fidelity.

- **Aya-23-8B**, multilingual instruction-tuned model, designed for cross-lingual tasks.
- **LLaMA-3.1-8B-Instruct**, a chat-aligned model with strict system–user templates.
- **BLOOMZ-7B1-mt**, multilingual instruction-tuned model by BigScience.

Prompting formats. Two prompting styles were employed across models to ensure consistency and reproducibility.

General format

Answer the following question in [LANGUAGE] clearly and concisely. Question: {question} Answer:

This general instruction-based template was used for all causal and instruction-tuned models (Mistral, OpenHathi, Qwen, Yi, Gemma, Airavata, Aya, LLaMA, BLOOMZ). It explicitly enforces the target [LANGUAGE] and promotes concise answers.

GPT-4.1 (chat format)

"role": "user", "content": "Question: {question}"

GPT-4.1 requires a chat-style JSON format with explicit user roles. This reflects its native API design, which allows role-based conversation management.

Table 6: Illustrative GPT-4.1 examples with Question, References (Indic & EN), Outputs (Indic & EN), and scores (chrF++/BLEU/BERT/LRMGS).

Language: Assamese
Question (EN): Khandesh region was captured by Mughal Emperor Akbar in which year?
Reference (Indic): ১৬০১ চনত মোগল সম্ৰাট আকবৰে খানদেশ অঞ্চল দখল কৰে।
Reference (EN): In 1601, the Mughal Emperor Akbar conquered the Khandesh region.
Output (Indic): খানদেশ অঞ্চল ১৬০১ চনত মোগল সম্ৰাট আকবৰে দখল কৰিছিল।
Output (EN): The Khandesh region was conquered by the Mughal Emperor Akbar in 1601.
Scores: chrF++ = 0.800 BLEU = 0.531 BERT = 0.970 LRMGS = 0.930

Language: Assamese
Question (EN): Which Anglo-Maratha war resulted in merging the Peshwas' territories with the Bombay Presidency?
Reference (Indic): তৃতীয় এংলো-মাৰাঠা যুদ্ধৰ ফলত পেশ্বৰসকলৰ ভুখণ্ড বম্বে ৰাষ্ট্ৰপতিৰ সৈতে একত্ৰিত হয়।
Reference (EN): The Third Anglo-Maratha War united the territory of the Peshwas with the Bombay Presidency.
Output (Indic): তৃতীয় এংলো-মাৰাঠা যুদ্ধৰ ফলত পেশ্বৱাৰ ভুখণ্ড বম্বে ৰাষ্ট্ৰপতিৰ সৈতে একত্ৰিত হৈছিল।
Output (EN): The Third Anglo-Maratha War united the Peshwa territory with the Bombay Presidency.
Scores: chrF++ = 0.883 BLEU = 0.588 BERT = 0.981 LRMGS = 0.906

Language: Dogri
Question (EN): The words "Satyameva Jayate" in the State Emblem adopted by the Government of India have been taken from which Upanishad?
Reference (Indic): "सत्यमेव जयते" शब्द मुंडक उपनिषद थमां लैते गेदे न।
Reference (EN): The words "Satyameva Jayate" are taken from the Mundaka Upanishad.
Output (Indic): "सत्यमेव जयते" शब्द मुण्डक उपनिषद थमां लैते गेदे न।
Output (EN): The words "Satyameva Jayate" are taken from the Mundaka Upanishad.
Scores: chrF++ = 0.905 BLEU = 0.597 BERT = 0.997 LRMGS = 0.905

Language: Dogri
Question (EN): In which caves is the Kailasha temple located?
Reference (Indic): कैलाशा मंदिर एलोरा गुफाएं च स्थित ऐ।
Reference (EN): Kailasha Temple is located in the Ellora Caves.
Output (Indic): कैलाश मंदिर एलोरा गुफाएं च स्थित ऐ।
Output (EN): Kailash Temple is located in Ellora Caves.
Scores: chrF++ = 0.875 BLEU = 0.809 BERT = 0.990 LRMGS = 0.896

Language: Maithili
Question (EN): What is FERA and when was it introduced?
Reference (Indic): 1973 कें विदेशी मुद्रा नियमन अधिनियम (फेरा) विदेशी मुद्रा, प्रतिभूतियक, मुद्रा कें आयात आ निर्यात आ विदेशीक कें द्वारा अचल संपत्ति कें अधिग्रहण सं संबंधित किछु भुगतान कें नियंत्रित करय कें लेल एकटा अधिनियम छै.
Reference (EN): The Foreign Exchange Regulation Act of 1973 (FERA) regulates foreign exchange, securities, currency, and certain payments related to acquisition of immovable property by foreigners.
Output (Indic): फेरा एकटा धार्मिक अनुष्ठान छी जे सिख धर्म में विवाहक समय होइत अछि। ई गुरु ग्रंथ साहिब केर समक्ष वर-वधू द्वारा चारि बेर घुमि-घुमि क' कएल जाइत अछि। फेरा सिख धर्म में गुरु नानक देव जी केर समय सँ शुरू भेल।
Output (EN): Fera is a religious ceremony in Sikhism performed during marriage, where the couple circles the Guru Granth Sahib four times; it dates back to the time of Guru Nanak Dev Ji.
Scores: chrF++ = 0.157 BLEU = 0.012 BERT = 0.826 LRMGS = 0.018

Language: Manipuri
Question (EN): When was the Employment Guarantee Scheme (EGS) introduced in Maharashtra?
Reference (Indic): ꯏꯃꯞꯂꯣꯏꯃꯦꯟꯇ ꯒꯦꯔꯦꯟꯇꯤ ꯁ꯭ꯀꯤꯝ (ꯏ.ꯖꯤ.ꯑꯦꯁ.) ꯑꯁꯤ ꯃꯥꯔꯆ ꯲꯸, ꯱꯹꯷꯲ꯗꯥ ꯍꯧꯗꯣꯀꯈꯤ॥
Reference (EN): The Employment Guarantee Scheme (EGS) was introduced on 28 March.
Output (Indic): ꯏ.ꯖꯤ.ꯑꯦꯁ. ꯁꯤꯝꯕꯤꯗꯨ ꯍꯧꯗꯣꯛꯈꯤꯕꯅꯤ ꯑꯃꯁꯤ ꯃꯍꯥꯔꯥꯁ꯭ꯠꯔꯗ ꯏꯃꯞ
Output (EN): E.G.S.—when did you start the scheme in Maharashtra?
Scores: chrF++ = 0.304 BLEU = 0.268 BERT = 0.949 LRMGS = 0.017