Revisiting Evaluation of Question Answering Systems in Low-Resource Indic Languages: Bridging Human and Metric Alignment

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

Evaluating Question Answering (QA) systems in low-resource Indic languages remains challenging due to the scarcity of annotated data, high linguistic diversity, and the absence of reliable evaluation metrics. Many Indian languages are severely underrepresented, making it difficult to accurately assess the performance of Large Language Models (LLMs) on QA tasks. Commonly used metrics like BLEU, ROUGE-L, and BERTScore, while successful in machine translation and resource-rich scenarios, tend to perform poorly in low-resource QA settings. These metrics often exhibit issues such as compressed scoring ranges, excessive zero scores, and weak alignment with human judgments. To overcome these limitations, this work introduces the LRM^{2}QAS (Language Robust Multi-aspect Metrics for Question Answering Systems). This composite evaluation framework integrates semantic similarity, factual completeness, numerical accuracy, and contextual relevance. The proposed metric is evaluated across eight Indic-language QA tasks using multiple LLMs, as well as on open-domain benchmarks NaturalQuestions (NQ) and TriviaQA (TQ). Across all settings, LRM^{2}QAS demonstrates stronger agreement with human evaluation, as measured by Pearson, Spearman, and Kendall correlation coefficients. Experimental findings highlight that LRM^{2}QAS provides more precise distinctions between model outputs and aligns more closely with human judgment, offering a reliable framework for evaluating multilingual QA in low-resource Indic languages.

1 Introduction

The development of Natural Language Processing (NLP) systems continues to face a fundamental challenge: many languages remain underrepresented due to the limited availability of annotated corpora and reliable evaluation benchmarks. This issue is particularly evident in linguistically rich regions such as India, where numerous languages with millions of speakers are still inadequately supported. Large Language Models (LLMs) offer a promising approach to bridging this gap by enabling knowledge transfer 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 on tasks such as summarisation (Pu et al., 2023; Goyal et al., 2023) and question answering (Zhao et al., 2023). However, their training and evaluation pipelines remain predominantly English-centric, which limits their ability to generalise across diverse linguistic contexts (Lai et al., 2023; Zhang et al., 2023; Ahuja et al., 2023) and highlights performance disparities between proprietary and open-source models (Ahuja et al., 2024).

Existing benchmarks are largely focused on understanding tasks and provide limited support for assessing generative outputs (Lai et al., 2023; Asai et al., 2023), and they also rely heavily on costly human annotations. LLM-based evaluation (Liu et al., 2023) offers a more scalable alternative; however, it introduces systematic biases, such as a preference for longer responses and outputs that resemble the model's own generations (Zheng et al., 2023; Shen et al., 2023). Consequently, even though multilingual pretraining has expanded generation capabilities across languages (Jiang et al., 2024), evaluation frameworks have not kept pace.

BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which remain widely used in Indic QA evaluation, are primarily based on surface-level overlap, limiting their ability to capture true answer quality. BLEU focuses on n-gram similarity, often penalising valid paraphrases while failing to account for factual consistency, whereas ROUGE emphasises recall and may reward outputs that include irrelevant or even hallucinated information. More recent embedding-based approaches, such as BERTScore (Zhang et al., 2020), attempt to address these short-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 703–718
July 2-7, 2026 ©2026 Association for Computational Linguistics

comings by incorporating contextual representations; however, they can still assign high scores to responses that appear semantically similar but are factually incorrect or insufficiently grounded. Likewise, chrF++ (Popović, 2017), which captures morphological variation through character-level comparisons, remains insensitive to deeper semantic meaning and contextual faithfulness. Despite these well-known limitations, evaluation in Indic QA settings continues to rely heavily on such automatic metrics, even as several datasets (Clark et al., 2020; Asai et al., 2021; Singh et al., 2025) have expanded coverage for low-resource languages. As a result, fluent yet factually inaccurate responses can still achieve high scores under commonly used evaluation frameworks, highlighting a critical gap in reliable assessment.

To address the lack of reliable evaluation methods for generative question answering in very low-resource Indic languages, this work builds on the (Rohera et al., 2024; Kumar et al., 2025) and introduces a newly curated human-annotated evaluation set covering Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi. The proposed LRM²QAS (Language-Robust Multi-aspect Metrics for Question Answering Systems) defines a unified evaluation framework that integrates pivot-based semantic similarity, nugget-level factual coverage, numeric fidelity, and evidence faithfulness, enabling assessment beyond surface-level lexical overlap. Experiments conducted across ten large language models demonstrate consistent discrimination of generative QA quality across languages and domains. To examine generalisability beyond Indic benchmarks, the framework is evaluated on standard open-domain QA datasets (Wang et al., 2023), including Natural Questions (NQ) and TriviaQA (TQ), under the same human evaluation protocol. The evaluation is grounded in human judgment, comprising 800 manually verified scores across five qualitative dimensions, with correlation analysis indicating stronger alignment with human judgments compared to existing metrics.

The contributions of this work are summarised as follows:

- Development of a human-aligned evaluation framework, LRM²QAS, for assessing generative QA systems across multiple qualitative dimensions.
- Construction of a human-annotated evaluation dataset for multiple very low-resource Indic languages, enabling reliable assessment in underrepresented settings.
- Extension of evaluation to existing human-annotated open-domain QA datasets, demonstrating the generalisability and robustness of the proposed framework.

2 Evaluation Protocol

2.1 Problem Definition

This work evaluates 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

Automatic evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and chrF++ (Popović, 2017) approximate answer quality primarily through lexical overlap or embedding similarity. Such approaches remain insufficient for factual question answering in very low-resource settings, where fluent yet factually incorrect responses may receive inflated scores. Independent reporting of multiple metrics further introduces ambiguity, as different dimensions may be implicitly prioritised without a principled aggregation strategy, reducing comparability across systems.

To address these limitations, **LRM^{2}QAS**[1] is defined as a structured multi-aspect evaluation framework that integrates semantic similarity, question-aware nugget coverage, numeric fidelity, and evidence faithfulness into a unified objective. The final score is computed using a multiplicative formulation:

$$\text{LRM}^2\text{QAS} = \prod_{\substack{k \in \{\text{BERT}, \\ \text{KC,NUM,EF}\}}} \text{EN-k}(R_{en}, \hat{A}_{en}, C_{en})^{\lambda_k}, \tag{1}$$

where λ_k denotes the weighting coefficient for each component. These parameters are not fixed and are adapted based on the evaluation setting. When human-annotated validation data is available,

[1] https://github.com/anuj0405/LRM2QAS.git

λ_k can be estimated by maximising alignment with human judgments. In the absence of such annotations, the coefficients are determined based on task characteristics and empirical observations.

In particular, answer length and content structure influence the relative importance of evaluation components. For open-domain QA datasets such as Natural Questions (NQ) and TriviaQA (TQ), answers are typically short, consisting of single tokens. In such cases, semantic similarity signals are less informative, and configurations such as $\lambda_{\text{BERT}} = 0.6$, $\lambda_{\text{KC}} = 0.9$, and $\lambda_{\text{NUM}} = \lambda_{\text{EF}} = 1$ prioritise factual coverage and correctness.

In contrast, Indic QA datasets contain longer, more descriptive answers with an average length of approximately 26 tokens. These responses require evaluation across multiple dimensions, including semantic alignment, factual completeness, and contextual grounding. Accordingly, configurations such as $\lambda_{\text{BERT}} = 0.9$, $\lambda_{\text{KC}} = 0.8$, and $\lambda_{\text{NUM}} = \lambda_{\text{EF}} = 1$ maintain a more balanced contribution across components. This design ensures that the evaluation remains robust across datasets with varying answer lengths and structural properties.

The multiplicative structure regulates interactions between evaluation components by limiting compensatory effects. Low scores in any dimension directly influence the final result, preventing weaker aspects from being masked by stronger ones. Unlike linear aggregation, which averages conflicting signals, this formulation enforces balanced performance across semantic, factual, numeric, and grounding dimensions, resulting in a more stable and human-aligned evaluation.

Notation. Each QA instance is represented as (Q, R), where Q denotes the Indic-language question, R the reference answer, and $\hat{A}$ the system-generated answer. English translations of Q and R are provided, and $\hat{A}$ is translated using IndicTrans2 (Gala et al., 2023) to enable evaluation within a shared representation space. The translated forms are denoted as Q_{en}, R_{en}, and $\hat{A}_{en}$. The grounding context is represented as $C_{en} = \{c_1, \ldots, c_m\}$.

Semantic similarity (EN-BERT).

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

Token-level embeddings $\mathbf{r}_i$ and $\mathbf{a}_j$ are extracted using RoBERTa-large (Zhang et al., 2020). This component captures semantic alignment between reference and generated answers.

Question-aware nugget coverage (EN-KC).

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

where $\mathbf{k}_i$ and $\mathbf{q}$ denote embeddings of clause c_i (segmented from R_{en}) and the question Q_{en}, respectively. Nuggets correspond to factual clauses segmented from R_{en}. Attention weights prioritise question-relevant content, ensuring that evaluation reflects factual coverage aligned with the query.

Numeric fidelity (EN-NUM).

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

where N_R and $N_{\hat{A}}$ denote the sets of numeric values extracted from the reference and generated answers using regular expressions. This component enforces consistency of quantitative information.

Evidence faithfulness (EN-EF).

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

where $\mathbf{a}$ denotes the sentence-level embedding of the generated answer and $\mathbf{c}$ denotes the embedding of a context sentence in C_{en}. The construction of C_{en} depends on dataset availability. For Indic QA settings, C_{en} is derived from translated question context. For datasets such as Natural Questions and TriviaQA, where explicit evidence passages are not available in the evaluation files, C_{en} is constructed from sentence-level segments of the reference answer. In this setting, EN-EF measures semantic alignment with reference content rather than external evidence attribution. Furthermore, the algorithm for LRM^{2}QAS is provided in the Appendix E.

3 Experiment Setup

This section presents the datasets used in this study, followed by the experimental setup, including data preparation and evaluation procedures.

Human Evaluation Protocol Human evaluation is conducted to assess alignment between automatic metrics and human judgment using 800 question–answer pairs, comprising 20 samples per language across eight Indic languages and five large

language models (GPT-4.1, Aya-23-8B, OpenHathi-7B-Hi-Base, Llama-3.1-8B-Instruct, and Gemma-2-9B-it). Two independent annotators evaluate responses using a structured 1–5 scale across five dimensions: *Factual Accuracy*, *Relevance*, *Clarity*, *Language Consistency*, and *Conciseness*. All annotations are manually verified to ensure fairness and cross-lingual consistency. Annotators are proficient in Maithili, Dogri, Sanskrit, Hindi, and Konkani, while outputs in Assamese, Manipuri, and Sindhi are translated into English and compared against reference answers. Each model output is evaluated using this unified protocol, yielding a matrix of human scores across all evaluation dimensions. The more detailed annotation guidelines and demography of annotators are provided in Appendix A.

An overall human score is computed as a weighted aggregation of the five dimensions, assigning higher importance to *Factual Accuracy* and *Relevance*. The annotation process is conducted by two annotators under the supervision of a domain expert over a four-month period, with disagreements occurring in fewer than 10% of samples and resolved through discussion and consensus. The finalised scores serve as ground truth for correlation analysis with automatic metrics using Pearson's r, Spearman's ρ, and Kendall's τ, as reported in Tables 1.

For open-domain datasets (NQ and TQ), evaluation is conducted using the provided question, reference answer, model predictions, and corresponding human correctness labels. Each model-generated response is treated as an independent evaluation instance, paired with its associated binary correctness label as the supervision signal. At the same time, reference answers are used to compute automatic metric scores. As these datasets do not include multi-dimensional annotations or explicit evidence passages, evaluation is restricted to correctness-based supervision and reference-based comparison. This design enables consistent evaluation across datasets with differing annotation granularity while maintaining comparability of metric behaviour. Correlation results with human judgments are reported in Table 2, and the overall evaluation pipeline is illustrated in Figure 1 (Appendix B). Additional implementation details and hyperparameter configurations are provided in Appendix C.

Metric	Pearson	Spearman	Kendall
DEEPSEEK-llm-7b-chat	0.282	0.247	0.176
BLEU	0.301	0.322	0.227
ROUGE_L	0.370	0.372	0.261
BERTScore	0.407	0.401	0.290
chrF++	0.420	0.40	0.30
LRM^{2}QAS	**0.494**	**0.511**	**0.395**

Table 1: Correlation between automatic evaluation metrics and human judgments across QA outputs.

Datasets. The evaluation is conducted on both multilingual Indic QA and open-domain QA benchmarks. The Indic dataset (Rohera et al., 2024) consists of 800 human-annotated QA instances spanning eight low-resource languages: Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi. This dataset is designed to capture evaluation challenges in low-resource multilingual settings, with detailed human judgments across multiple qualitative dimensions.

To assess generalisability, open-domain QA datasets, Natural Questions (NQ) and TriviaQA (TQ), are also used (Wang et al., 2023). These datasets differ in structure and annotation scheme, providing binary human correctness labels and typically shorter answers. Their inclusion enables evaluation under diverse conditions, including variation in answer length, domain, and annotation granularity. Additional dataset details are provided in Appendix B.

4 Results

As shown in Table 1, **LRM^{2}QAS** achieves the highest correlation with human judgments among all evaluated metrics, with approx Pearson 0.49, Spearman 0.51, and Kendall 0.40, exceeding lexical metrics such as BLEU and ROUGE-L, embedding-based metrics including BERTScore and chrF++, and the LLM-based evaluator DEEPSEEK-llm-7b-chat. Consistent trends are observed in open-domain QA settings Table 2. For NaturalQuestions (NQ) and TriviaQA (TQ), human judg-

Dataset	Metric	Pearson	Spearman	Kendall
NQ	BLEU	0.166	0.416	0.356
	ROUGE_L	0.293	0.463	0.355
	BERTScore	0.314	0.393	0.321
	chrF++	0.392	0.468	0.394
	LRM^{2}QAS	**0.443**	**0.493**	**0.402**
TQ	BLEU	0.085	0.230	0.197
	ROUGE_L	0.181	0.342	0.303
	BERTScore	0.137	0.158	0.129
	chrF++	0.182	0.238	0.195
	LRM^{2}QAS	**0.384**	**0.415**	**0.339**

Table 2: Correlation between automatic evaluation metrics and human judgments on open-domain QA benchmarks: NaturalQuestions (NQ) and TriviaQA (TQ).

	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
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: LRM^{2}QAS scores across eight Indic languages for different LLMs. Scores are averaged over all samples for each language using English-based prompts. GPT-4.1 consistently outperforms other models across all languages.

ments are provided in a binary correctness setting, and LRM^{2}QAS exhibits a higher correlation with human evaluation than both lexical metrics and BERTScore, which relies on embedding-based semantic similarity, a component similar to LRM^{2}QAS. On NQ, LRM^{2}QAS achieves Pearson 0.44, Spearman 0.49, and Kendall 0.40, while on TQ it attains Pearson 0.38, Spearman 0.41, and Kendall 0.34. These results indicate that integrating semantic similarity with nugget-level factual coverage, numeric fidelity, and evidence faithfulness yields more reliable alignment with human judgment across both low-resource Indic and open-domain QA benchmarks.

Building on this validation, Table 3 applies LRM^{2}QAS to compare LLM performance across eight Indic languages. The results show clear variation across models and languages, with open-source model averages ranging from approximately 0.14 (OpenHathi-7B-Hi-Base) to 0.30 (LLaMA-3.1-8B-Instruct), reflecting the impact of instruction tuning and architecture. Recent instruction-tuned models such as LLaMA-3.1-8B-Instruct and Gemma-2-9B-it consistently outperform earlier baselines, while extremely low-resource languages such as Manipuri remain challenging (scores as low as 0.04). GPT-4.1 achieves the highest scores across all languages (up to 0.43 on Hindi), highlighting the gap between proprietary and open-source systems. LRM^{2}QAS captures these fine-grained cross-lingual differences more clearly than conventional metrics. Further model configurations of LLMs, prompts and visualisation plots are detailed in Appendix D and F.

Diagnostic analyses are conducted to characterise the behaviour of LRM^{2}QAS beyond aggregate correlation measures. The ablation study quantifies metric–human divergence using absolute error analysis, examines stability as a function of answer length (token count), and evaluates cross-lingual robustness under translation-based English pivoting. The results reveal stronger alignment between LRM^{2}QAS and human judgments for short, fact-oriented responses, with a systematic increase in absolute error observed for longer or more stylistically complex outputs. Component-level analysis reveals that question-aware nugget coverage and evidence faithfulness significantly contribute to metric–human agreement, whereas divergence in longer answers arises when discourse-level completeness and elaboration, as reflected in human ratings, extend beyond the factual and grounding signals emphasised by LRM^{2}QAS. Cross-lingual analyses demonstrate stable relative system rankings across languages despite variation in absolute scores, indicating that translation-induced paraphrastic variation primarily affects score magnitude rather than relative ordering. Additional diagnostic details are provided in Appendix G.

5 Conclusion

The proposed LRM^{2}QAS metric introduces a language-robust, multi-aspect evaluation framework that jointly captures semantic similarity, nugget-level factual coverage, numeric fidelity, and evidence faithfulness to assess QA quality. Experiments across eight low-resource Indic languages and open-domain QA benchmarks demonstrate consistently stronger alignment with human judgments compared to existing metrics. Findings show that LRM^{2}QAS is reliable for evaluation in both multilingual and open-domain settings, providing a principled foundation for benchmarking generative QA systems. Furthermore, future work will explore systematic strategies for learning λ configurations tailored to specific languages and domains, enabling more effective and reusable deployment across diverse evaluation settings.

Limitations

The proposed evaluation framework has several limitations that constrain its scope. LRM^{2}QAS relies on translating system-generated answers into English using IndicTrans2 to obtain a shared representation space, which may introduce paraphrastic variation or minor semantic drift, particularly for ultra-low-resource languages with weaker machine translation support. Although cross-lingual analyses indicate stable relative system rankings, absolute score values may still be affected by translation quality. Furthermore, the evaluation is limited to eight very low-resource Indic languages due to dataset availability and annotator coverage.

In addition, LRM^{2}QAS is primarily designed for short, factual question answering. Diagnostic analyses indicate reduced alignment with human judgments for longer or stylistically complex answers, where discourse-level completeness and elaboration contribute more strongly to human evaluation than nugget-based scoring. The selection of the λ parameters also introduces a limitation, as these values are currently chosen based on intuitive reasoning and validated through empirical tuning (e.g., grid search) on the available datasets. While such configurations perform well for the considered Indic QA and open-domain benchmarks, they may not generalise consistently across different languages, domains, or tasks. Consequently, the current formulation may not directly extend to tasks such as summarisation, which involve longer outputs, discourse structure, and content selection beyond discrete factual units, and would require task-specific adaptations.

References

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.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Sarvam AI. 2023. Openhathi-7b-hi-v0.1-base. https://huggingface.co/sarvamai/OpenHathi-7B-Hi-v0.1-Base. Based on LLaMA2 for Hindi/English/Hinglish support.

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.

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*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint*. ArXiv:2401.15006.

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, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*. ArXiv:2310.06825.

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.

Anuj Kumar, Satyadev Ahlawat, Yamuna Prasad, and Virendra Singh. 2025. LRMGS: A language-robust metric for evaluating question answering in very low-resource Indic languages. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Cohere Labs. 2025. Aya-23-8b. https://huggingface.co/CohereLabs/aya-23-8B. Open weights research release with multilingual instruction tuning.

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.

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.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint*. ArXiv:2403.08295.

Unsloth AI / Qwen Team. 2025. Qwen2.5-7b instruct. https://huggingface.co/unsloth/Qwen2.5-7B-Instruct. Instruction-tuned 7.61B multilingual model supporting long context.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-QA evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

BigScience / xP3mt Team. 2023. Bloomz-7b1-mt. https://huggingface.co/bigscience/bloomz-7b1-mt. Multilingual instruction-following model finetuned on xP3mt.

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 Human Evaluation

Ethical Considerations This study focuses on advancing multilingual question-answering (QA) evaluation for low-resource Indic languages through the development and validation of the Language-Robust Metric for Generative QA Systems (LRM^{2}QAS). All experiments and human evaluations were carried out exclusively for academic research purposes. The dataset used in this work consists of publicly available factual QA content and does not contain any personally identifiable or culturally sensitive information. The human evaluation process was designed to ensure fairness, transparency, and linguistic neutrality across all languages. Annotators were instructed to maintain objectivity, avoid regional or cultural bias, and verify factual correctness before assigning final ratings. All data and annotations produced through this process are intended solely for research and model evaluation purposes and should not be reused in downstream applications without proper ethical oversight.

Annotator Demographics and Treatment Two annotators participated in the human evaluation phase. They possess prior experience in multilingual NLP and human evaluation methodologies. The annotators are proficient in Maithili, Dogri, Sanskrit, Hindi, and Konkani, while outputs in Assamese, Manipuri, and Sindhi were translated into English for assessment. Both annotator participation was voluntary, they underwent a brief orientation to ensure a consistent understanding of rating criteria, including factual accuracy, relevance, and clarity. The task involved no sensitive or distressing material, and care was taken to maintain the annotator's well-being throughout the process. The further evaluation pipeline is illustrated in Figure 1.

B Dataset

This study uses two complementary datasets to evaluate generative QA systems: (i) a newly curated human-annotated multilingual Indic QA dataset designed for very low-resource settings, and (ii) standard open-domain QA benchmarks (NaturalQuestions and TriviaQA) used to assess cross-domain generalisation. Together, these datasets enable analysis of metric behaviour across languages, resource conditions, and evaluation regimes.

Indic Multilingual QA Dataset. The primary dataset (Rohera et al., 2024) consists of 800 human-annotated QA instances spanning eight low-resource Indic languages: Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi. Each language contributes 200 QA pairs, with system outputs generated by multiple large language models. The dataset is specifically used to study evaluation challenges in low-resource multilingual factual QA, where reliably assessing model outputs remains difficult.

Human Annotation and Scoring. Human evaluation is conducted to assess alignment between automatic metrics and human judgment using 800 question–answer pairs, comprising 20 samples per language across eight Indic languages and five large language models (GPT-4.1, Aya-23-8B, OpenHathi-7B-Hi-Base, Llama-3.1-8B-Instruct, and Gemma-2-9B-it). Two independent annotators evaluate responses using a structured 1–5 scale across five dimensions: *Factual Accuracy*, *Relevance*, *Clarity*, *Language Consistency*, and *Conciseness*. All annotations are manually verified to ensure fairness and cross-lingual consistency. Annotators are proficient in Maithili, Dogri, Sanskrit, Hindi, and Konkani, while outputs in Assamese, Manipuri, and Sindhi are translated into English and compared against reference answers. Each model output is evaluated using this unified protocol, yielding a matrix of human scores across all evaluation dimensions. A diagram to so show the human evaluation pipeline is provided in Figure 1.

An overall human score is computed as a weighted aggregation, prioritising factual correctness (30%) and relevance (25%), followed by clarity (20%), language consistency (15%), and conciseness (10%). Each record includes model outputs, verified human ratings, and corresponding automatic metric scores, enabling systematic meta-evaluation and correlation analysis.

Example-Based Human Evaluation. Table 5 presents representative GPT-4.1 examples from Maithili, Hindi, and Assamese, evaluated across five human rating dimensions: *Factual Accuracy (FA)*, *Relevance (RE)*, *Clarity (CL)*, *Language Consistency (LA)*, and *Conciseness (CO)*. Each model response was rated on a 1–5 scale, and the overall *Human Score (HS)* was computed through a

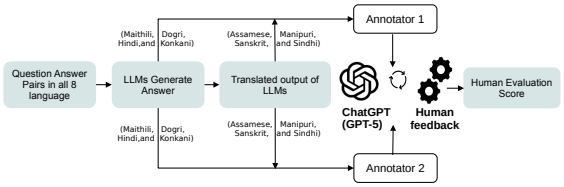

Figure 1: Human evaluation pipeline integrating annotator review and ChatGPT-5–assisted scoring for multilingual QA across eight Indic languages.

weighted aggregation that prioritises factual correctness and task relevance. The normalised formulation is given as:

$$\text{HS} = 100 \times \frac{\sum_i w_i (s_i - 1)}{40}, \quad w_i = [3, 2.5, 2, 1.5, 1],$$

where s_i denotes the individual score for each criterion, and w_i represents its corresponding weight in the aggregation.

This formulation maps the 1–5 scale to a 0–100 range, where higher HS values indicate stronger factual precision and linguistic consistency. As observed in Table 5, the Maithili question on *Elasticity* attains a high HS (83.75) due to factual and linguistic alignment, while the Maithili example on election symbols records a low HS (17.5) for factual hallucination. The Hindi question on "Detroit of India" yields a similar low HS (12.5) due to incorrect factual content, whereas the Assamese example about *Hamlet's father's death* achieves a near-perfect HS (98), reflecting complete factual and semantic correctness. These examples demonstrate that human evaluation captures deeper dimensions of factual reliability and clarity than surface-level automatic metrics, providing a grounded benchmark for assessing the quality of multilingual QA.

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

Dataset Size and Length Characteristics. Table 4 summarises per-language statistics of the Indic QA dataset. Each language contains 200 QA pairs. The average question length ranges from 7.56 tokens (Sanskrit) to 11.22 tokens (Dogri), while average answer length ranges from 19.32 tokens (Sanskrit) to 30.50 tokens (Dogri and Sindhi). Across all languages, answers are consistently longer than questions, typically by a factor of 2.5×–3×, reflecting the nature of factual QA where concise questions often require multi-clause responses.

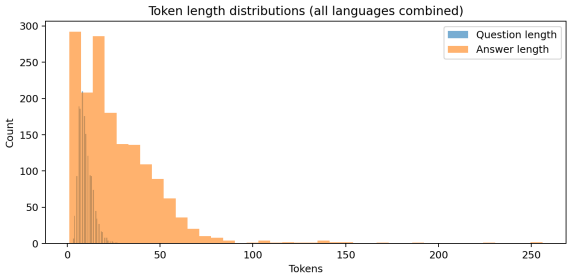

Figure 2: Token-length distributions aggregated across all languages. Questions are short and tightly clustered, while answers are longer and exhibit a right-skewed distribution with a long tail.

Figure 2 further illustrates these trends by showing the aggregated token-length distributions. Question lengths are concentrated within a narrow range (approximately 5–15 tokens), indicating low variance in prompt size. In contrast, answer lengths display a broader distribution with a pronounced right skew, occasionally exceeding 200 tokens. This clear separation between question and answer lengths reduces confounding effects during evaluation. It highlights the need for evaluation mechanisms that capture partial factual coverage and clause-level correctness in longer responses.

Open-domain QA datasets. To examine generalisability beyond Indic benchmarks, NaturalQuestions (NQ) and TriviaQA (TQ) are also used (Wang et al., 2023). These datasets provide open-domain QA instances with binary human correctness annotations and differ substantially from the Indic dataset in terms of language, domain, and answer style. Their inclusion allows analysis of metric robustness under simpler human judgment schemes and more diverse factual content.

C Experimental Setup and Metrics

Experiments are conducted on eight low-resource Indic languages: Assamese, Dogri, Hindi, Konkani, Maithili, Manipuri, Sanskrit, and Sindhi, using 200 question–answer pairs per language. For each

(Q, R) pair, model-generated answers $\hat{A}$ are evaluated using both automatic metrics and human-annotated scores. Evaluation is performed in an inference-only setting, and results are reported per language and per system, followed by correlation analysis with human judgments using Pearson's r, Spearman's ρ, and Kendall's τ.

Generation Protocol. Model outputs are generated using a structured chat-based prompting strategy. Each input is formatted using a system–user template, where the system instruction enforces generation strictly in the target language using its native script, and the user prompt provides the question along with a directive for concise answering. This design ensures consistent multilingual generation and reduces unintended language mixing. All models are used without any parameter updates or fine-tuning.

Inference Settings. All model outputs used in the evaluation are generated under a consistent inference configuration. Generation is performed using the HuggingFace $\mathrm{transformers}$ pipeline in $\mathrm{float16}$ precision across 2 NVIDIA V100 GPUs, with each model executed on a dedicated device. Decoding follows a deterministic greedy search ($\mathrm{do\_sample}$=false) with a maximum of 64 new tokens, ensuring stable, comparable outputs across systems. Tokenisers use left padding for decoder-only models to maintain consistent behaviour during generation. Inference is performed sequentially without gradient computation.

Metric Configuration. Standard automatic metrics are computed using widely adopted configurations. BLEU is computed using SacreBLEU at the corpus level. ROUGE-L is reported as the F_1 score for the longest common subsequence. chrF++ is computed with $\beta = 2$ to better capture morphological variation. BERTScore is computed using the $\mathrm{roberta\text{-}large}$ encoder with the F_1 aggregation variant, without baseline rescaling. All evaluations are performed in a unified English-projected space to ensure comparability across languages.

LRM^{2}QAS Configuration. The proposed LRM^{2}QAS metric combines semantic similarity, question-aware nugget coverage, numeric fidelity, and evidence faithfulness through a multiplicative formulation. The weighting parameters λ_k are selected based on dataset characteristics rather than fixed globally. For the Indic QA dataset, where answers are longer and more descriptive, a balanced configuration ($\lambda_{\text{BERT}} = 0.9$, $\lambda_{\text{KC}} = 0.8$) is used. For open-domain datasets such as NQ and TQ, where answers are shorter and often factoid in nature, the configuration ($\lambda_{\text{BERT}} = 0.6$, $\lambda_{\text{KC}} = 0.9$) prioritises factual coverage.

Nugget extraction is performed using clause-level segmentation of the reference answer, followed by question-aware attention weighting based on cosine similarity in a normalised embedding space. Numeric fidelity is computed using the overlap between extracted numeric expressions. Evidence of faithfulness is approximated through sentence-level semantic similarity between the generated answer and reference-aligned context. All embedding-based components use the $\mathrm{all\text{-}MiniLM\text{-}L6\text{-}v2}$ sentence encoder with normalised embeddings. The λ parameters are selected through empirical validation, including grid search over candidate values, to ensure stable alignment with human judgments.

LLM-based Evaluation. In addition to human evaluation, model-generated responses are assessed using LLM-based raters to examine consistency with human judgments. Specifically, $\mathrm{deepseek\text{-}ai/DeepSeek\text{-}LLM\text{-}7B\text{-}Chat}$ and $\mathrm{deepseek\text{-}ai/DeepSeek\text{-}V2\text{-}Lite\text{-}Chat}$ are used as structured evaluators, where each $(Q, R, \hat{A})$ instance is scored across five qualitative dimensions: *Factual Accuracy*, *Relevance*, *Clarity*, *Language Consistency*, and *Conciseness*. Evaluation is performed using a constrained prompt that enforces JSON-formatted outputs, and decoding is carried out deterministically ($\mathrm{do\_sample}$=false) to ensure consistency. The resulting scores, originally on a 1–5 scale, are normalised to the $[0, 100]$ range for direct comparison with human annotations. These LLM-based scores are used for auxiliary analysis and correlation comparison, complementing but not replacing human evaluation.

D Evaluated Systems and Prompting Setup

A diverse set of large language models (LLMs) is used to benchmark performance across low-resource Indic languages. The selection includes both Indic-focused models and general-purpose multilingual systems, enabling comparison across varying levels of linguistic specialisation. All models are evaluated under a unified prompting and inference framework to ensure consistency and reproducibility.

Evaluated Models. The evaluation includes the following models:

- **Mistral-7B** (Jiang et al., 2023), a decoder-only causal model.
- **OpenHathi-7B-Hi-Base** (AI, 2023), designed for Hindi and related Indic languages.
- **Qwen2.5-7B-Instruct** (Team, 2025), a multilingual instruction-tuned model.
- **Yi-1.5-9B-Chat** (AI et al., 2025), a chat-oriented decoder-only model.
- **GPT-4.1** (OpenAI et al., 2024), used as a high-capacity proprietary baseline.
- **Gemma-2-9B-it** (Team, 2024), an instruction-tuned chat model.
- **Airavata-7B** (Gala et al., 2024), an Indic-focused instruction model.
- **Aya-23-8B** (Labs, 2025), a multilingual model designed for cross-lingual tasks.
- **LLaMA-3.1-8B-Instruct**, a chat-aligned model with structured prompting.
- **BLOOMZ-7B1-mt** (xP3mt Team, 2023), a multilingual instruction-tuned model.

Prompting Strategy. To ensure consistent generation across models and languages, a unified prompting approach is adopted. For instruction-tuned and causal models, a standard template is used to enforce language control and concise responses:

Instruction-based Prompt

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

For chat-based models such as GPT-4.1, prompts are adapted to their native conversational format while preserving the same instruction semantics:

Chat-style Prompt

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

This unified design ensures that all models receive equivalent task instructions, reducing variability due to prompt formulation while maintaining compatibility with model-specific interfaces.

LLM as Judge Prompt Design. For LLM-assisted scoring and structured evaluation, a consistent evaluation prompt is used to guide scoring behaviour across models and languages. The prompt enforces comparison between the generated answer, the reference answer, and the original question, with explicit emphasis on factual correctness:

Evaluation Prompt for LLM-assisted Scoring and Human Verification

Evaluate the quality of the model's response to a question by strictly comparing it with the reference answer and the original question.

Assign scores on a 1–5 scale (decimals allowed) for each criterion below:

Factual Accuracy: Degree to which the answer is correct with respect to the reference (5 = fully correct, 3 = partially correct, 1 = incorrect or hallucinated).

Relevance: The Extent to which the answer directly addresses the question (5 = fully relevant, 1 = off-topic).

Clarity: Readability, coherence, and logical structure of the response (5 = clear and well-structured, 1 = confusing).

Language Consistency: Ensure the language of the output matches the input question. Penalise mixed or incorrect languages.

Concise: Completeness without unnecessary verbosity (5 = concise and sufficient, 1 = verbose or incomplete).

Factual accuracy should dominate the judgment; fluent but incorrect answers must receive low scores.

Output format (JSON only):

```
{
  "Factual Accuracy": score,
  "Relevance": score,
  "Clarity": score,
  "Language Consistency": score,
  "Conciseness": score
}
```

E LRM^{2}QAS Algorithm

The proposed LRM^{2}QAS metric is defined as a structured composition of multiple evaluation signals designed to capture semantic, factual, numeric, and grounding aspects of QA quality. The formulation integrates semantic similarity (EN-BERT),

question-aware nugget coverage (EN-KC), numeric fidelity (EN-NUM), and evidence faithfulness (EN-EF), enabling consistent and reproducible evaluation across multilingual and low-resource settings. The computation proceeds through the following steps:

Algorithm 1: Computation of LRM^{2}QAS Metric (Symbolic Form)

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

Output: LRM^{2}QAS $\in [0, 1]$

Step 1: Preprocessing

Translate inputs using IndicTrans2:
$Q_{en}, R_{en}, \hat{A}_{en} \leftarrow \text{Trans}_{\text{IndicTrans2}}(Q, R, \hat{A})$; split question into sentence-level units $C_{en} = \{c_1, \ldots, c_m\}$.

Step 2: Semantic Similarity (EN-BERT)

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

Step 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

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

Step 4: Numeric Fidelity (EN-NUM)

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

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

Step 5: Evidence Faithfulness (EN-EF)

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

Step 6: Aggregation

$$\text{LRM}^2\text{QAS} = (\text{EN-BERT})^{\lambda_{\text{BERT}}} (\text{EN-KC})^{\lambda_{\text{KC}}} (\text{EN-NUM})^{\lambda_{\text{NUM}}} (\text{EN-EF})^{\lambda_{\text{EF}}}.$$

The multiplicative formulation ensures that low scores in any individual component reduce the final score, preventing compensation across dimensions and enforcing balanced evaluation across semantic similarity, factual coverage, numeric consistency, and grounding signals.

F Visualization Plots and Example Analysis

System-wise Metric Comparison. Figure 3 compares ten large language models across five automatic evaluation metrics averaged over eight Indic languages. As observed, **BERTScore** consistently yields the highest absolute values (≈ 0.80–0.88) across all systems, reflecting its embedding-based similarity with limited discriminative capacity. Lexical metrics such as **BLEU** (0.001–0.068) and **ROUGE-L** (0.001–0.054) remain low and tightly clustered, underscoring their weakness in measuring semantic adequacy in multilingual QA. **chrF++** achieves slightly higher differentiation (0.02–0.28), with GPT-4.1 performing best on lexical overlap. The proposed **LRM^{2}QAS** metric shows clear and interpretable separation among systems ranging from 0.05 (Yi-1.5-9B-Chat) to 0.41 (GPT-4.1), followed by LLaMA-3.1-8B-Instruct (0.33) and Qwen2.5-7B-Instruct (0.30). Gemma-2-9B-it (0.29), Mistral-7B (0.21), and Airavata-7B (0.24) exhibit moderate alignment, while BLOOMZ-7B1-mt and OpenHathi-7B-Hi-Base remain the lowest. Overall, LRM^{2}QAS demonstrates stronger discriminative power and better reflection of factual and semantic correctness than lexical or embedding-based baselines.

Example Analysis. Table 5 shows four representative GPT-4.1 QA examples across Maithili, Hindi, and Assamese, illustrating the relationship between human judgments and automatic metric scores. The first Maithili example on *Elasticity* demonstrates a strong factual and linguistic match, achieving high human ratings (overall 83.75) and corresponding strong automatic scores (LRM^{2}QAS: 0.787, chrF++: 0.365). In contrast, the second Maithili example about election symbols contains a factual hallucination ("black horse"), reflected in both a low human score (17.5) and degraded metric values (LRM^{2}QAS 0.100). The Hindi example, where "Bangalore" replaces "Chennai" as the correct answer, also receives a low human score (12.5), confirming the model's factual error despite fluent phrasing. Conversely, the Assamese example on *Hamlet's father's death* achieves near-perfect human ratings (overall 98), aligning with high semantic correctness and moderate metric performance (LRM^{2}QAS 0.495). These patterns indicate that human-evaluated factual precision and relevance strongly influence metric behaviour; faithful, grounded responses consistently yield higher scores across both human and automatic evaluations, whereas hallucinated or off-topic content is penalised by all metrics, especially LRM^{2}QAS.

G Ablation Study

This section presents a diagnostic evaluation of LRM^{2}QAS through error stratification, answer-length-conditioned stability analysis, and cross-lingual robustness assessment. The analysis is intended to characterise systematic divergence patterns between metric scores and human judgments, rather than to maximise absolute correlation values.

Table 5: Illustrative GPT-4.1 examples with Questions, References, Outputs, Human Evaluation Scores, and Automatic Metric Scores.

Language: Maithili **Question (EN):** Define elasticity. Explain its types and importance in economic decisions. **Reference (Indic):** लोच एकटा आर्थिक अवधारणा छै जेकर उपयोग कोनों वस्तु या सेवा केर मांग में बदलावक माप लेल होय छै जे मूल्यक परिवर्तनक साथ बदलय छै। कोनों उत्पाद के लोचदार कहल जायत छै जखन मूल्य बढ़ला या घटला पर मांग आनुपातिक सं अधिक बदलय छै। **Reference (EN):** Elasticity measures the change in demand for a good or service relative to price changes. A product is elastic if demand changes more than proportionally with price variations. **Output (Indic):** लोच, एगो आर्थिक अवधारणा हो जहाँ एक वस्तु या सेवा के मूल्य में परिवर्तन के साथ माँग में परिवर्तन की मात्रा दर्शाए जात होला। आर्थिक निर्णय में लोच महत्वपूर्ण होला कयोंकि... **Output (EN):** Elasticity represents the change in demand corresponding to price changes. It is important in economic decisions because it affects pricing and policy choices. **Human Scores:** Factual = 4, Relevance = 5, Clarity = 4, Language Consistency = 4.5, Conciseness = 4, Overall = 83.75 **Automatic Metrics:** BLEU = 0.134 ROUGE-L = 0.376 chrF++ = 0.365 LRM^{2}QAS = 0.787
Language: Maithili **Question (EN):** What were the election symbols of the Congress Party and Bharatiya Jana Sangh in the first general election of 1952? **Reference (Indic):** 1952 के पहिल आम चुनाव में कांग्रेस पार्टी के प्रतीक बैल के जोड़ी आ भारतीय जनसंघ के प्रतीक दीपक छल। **Reference (EN):** In the first general election of 1952, the Congress Party's symbol was a pair of oxen and Bharatiya Jana Sangh's was a lamp (Deepak). **Output (Indic):** कांग्रेस पार्टी ने अपना चुनावी प्रतीक के रूप में काला घोड़ा चुना था। **Output (EN):** The Congress Party chose a black horse as its election symbol. **Human Scores:** Factual = 1, Relevance = 1, Clarity = 2, Language Consistency = 5, Conciseness = 2, Overall = 17.5 **Automatic Metrics:** BLEU = 0.020 ROUGE-L = 0.256 chrF++ = 0.249 LRM^{2}QAS = 0.100
Language: Hindi **Question (EN):** Which city is called "Detroit of India" for its automobile industry? **Reference (Indic):** चेन्नई को "एशिया का डेट्रायट" या "भारत का डेट्रायट" कहा जाता है क्योंकि वहाँ प्रमुख ऑटोमोबाइल उद्योग स्थित हैं। **Reference (EN):** Chennai is nicknamed the "Detroit of India" due to the presence of major automobile manufacturing industries. **Output (Indic):** बेंगलुरु। **Output (EN):** Bangalore. **Human Scores:** Factual = 1, Relevance = 1, Clarity = 2, Language Consistency = 3, Conciseness = 2, Overall = 12.5 **Automatic Metrics:** BLEU = 0.024 ROUGE-L = 0.211 chrF++ = 0.213 LRM^{2}QAS = 0.096
Language: Assamese **Question (EN):** How did Hamlet's father die? **Reference (Indic):** মই তোমাৰ দেউতাকৰ আত্মা, ভূতে হেমলেটক কয়। ক্লডিয়াছে তেওঁৰ কাণত বিষ লগাই হত্যা কৰিছিল। **Reference (EN):** Hamlet's father was murdered by his brother Claudius, who poured poison into his ear while he slept. **Output (Indic):** হেমলেটৰ পিতৃক তেওঁৰ ভাই ক্লডিয়াছে বিষ মিশ্রিত পানীয় দি হত্যা কৰিছিল। **Output (EN):** Hamlet's father was killed by his brother Claudius with a drink laced with poison. **Human Scores:** Factual = 4.8, Relevance = 5, Clarity = 5, Language Consistency = 5, Conciseness = 4.8, Overall = 98 **Automatic Metrics:** BLEU = 0.008 ROUGE-L = 0.250 chrF++ = 0.187 LRM^{2}QAS = 0.495

G.1 Metric–Human Disagreement Analysis

To explicitly quantify divergence between LRM^{2}QAS and human evaluation, the absolute disagreement between the two scores is analysed. For each instance i, the absolute error is defined as:

$$\text{Error}_i = \left|\text{LRM}^2\text{QAS}_i - \text{HumanScore}_i\right|,$$

where HumanScore_i denotes the normalized aggregate human rating. The *Mean Error* reported in sub-

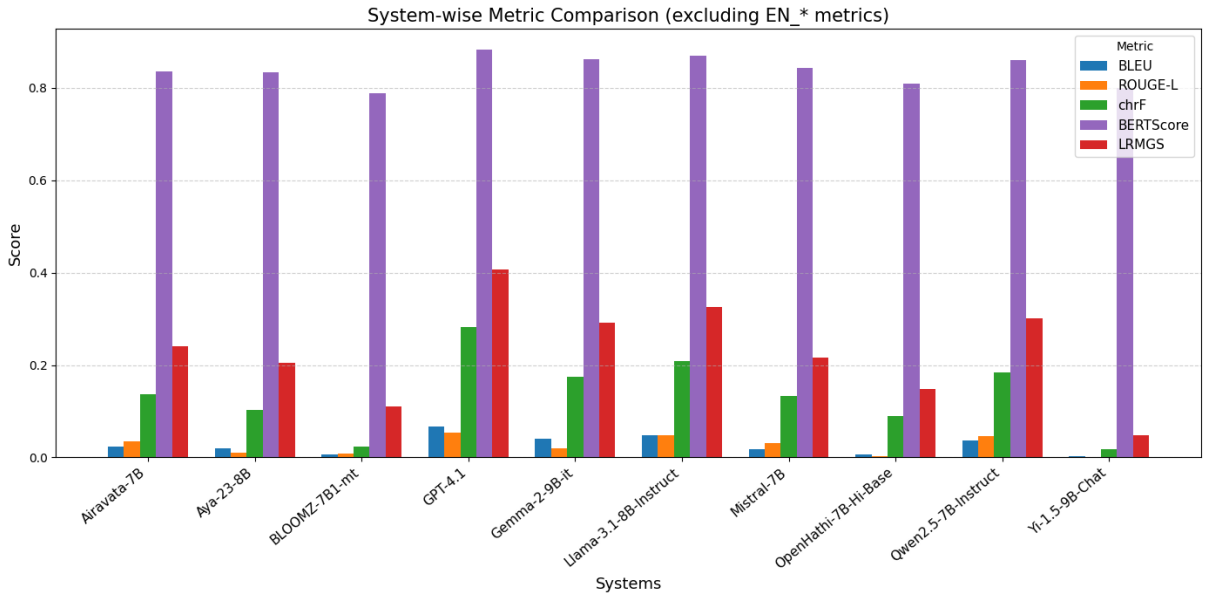

Figure 3: System-wise comparison of ten large language models across five automatic metrics averaged over eight Indic languages. LRM^{2}QAS exhibits the most consistent and discriminative pattern, effectively distinguishing higher-quality systems (GPT-4.1, LLaMA-3.1-8B-Instruct) from weaker ones (BLOOMZ-7B1-mt, Yi-1.5-9B-Chat), unlike traditional lexical metrics that remain compressed.

sequent tables corresponds to the arithmetic mean of Error$_i$ over all instances within a given subset and reflects the average magnitude of disagreement rather than rank inconsistency.

Instances are partitioned into three error regimes: low ($\leq$20), medium (20–50), and high ($\geq$50). Table 6 summarises the distribution of examples across these regimes. A substantial proportion of examples falls into the high-error bucket, indicating systematic disagreement in absolute score magnitude. Qualitative inspection reveals that these cases are frequently associated with longer or more elaborate answers, where discourse-level completeness, stylistic richness, or redundancy positively influence human judgments, while remaining largely unaccounted for by nugget-based scoring. Figure 4 visualises the prevalence of these disagreement regimes.

Table 6: LRM^{2}QAS–Human disagreement profile grouped by absolute error magnitude.

Error Bucket	Count	Mean Error	Mean LRM^{2}QAS	Mean Human
Low ($\leq$20)	31	0.16	0.08	0.17
Medium (20–50)	263	0.17	0.16	0.33
High ($\geq$50)	346	0.50	0.33	0.83

G.2 Role of Nugget Coverage (KC)

To isolate the contribution of question-aware nugget coverage, correlations between individual

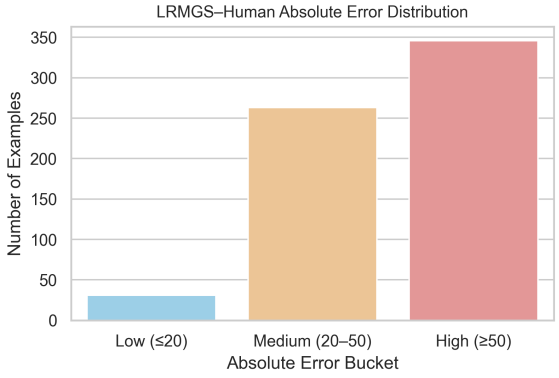

Figure 4: Distribution of absolute disagreement between LRM^{2}QAS and human judgments across error regimes.

LRM^{2}QAS components and human judgments are analysed. As shown in Table 7, the KC component exhibits strong agreement with human evaluation, achieving Pearson 0.49, Spearman 0.50, and Kendall 0.36, comparable to the full LRM^{2}QAS score and higher than standalone semantic similarity. This indicates that clause-level factual coverage captures aspects of answer quality that are not fully reflected by embedding-based similarity alone.

Further analysis conditions KC's behaviour on answer length. Table 8 shows that KC maintains moderate to strong correlation with human judgments across all length regimes, with the highest correlation observed for long answers (Pearson

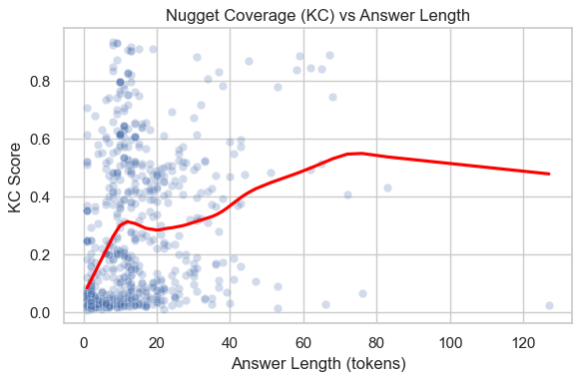

Figure 5: Relationship between question-aware nugget coverage (KC) and answer length, illustrating increasing coverage followed by saturation for longer answers.

Table 7: Correlation between individual LRM^{2}QAS components and human judgments.

Component	Pearson	Spearman	Kendall
BERTScore	0.46	0.45	0.32
KC (Nugget Coverage)	0.49	0.50	0.36
NUM (Numeric Fidelity)	0.07	0.08	0.07
EF (Evidence Faithfulness)	0.54	0.53	0.39
LRM^{2}QAS	**0.49**	**0.51**	**0.39**

0.57). Short answers also exhibit high correlation (0.52), suggesting that nugget matching remains informative even when answers are concise. Medium-length answers show comparatively lower correlation, reflecting increased variability in how additional content is structured and expressed.

Figure 5 visualises the relationship between answer length and KC score. KC increases with answer length up to a moderate range, reflecting improved factual coverage, and then saturates for longer answers. This saturation mirrors trends observed in human–metric disagreement analyses, indicating that nugget coverage prioritises core factual completeness rather than discourse-level elaboration.

G.3 LRM^{2}QAS under Answer Length

The stability of the full LRM^{2}QAS score is examined as a function of answer length, measured in tokens. Answers are partitioned into short, medium, and long bins based on empirical quantiles. For each bin, Table 9 reports the mean and standard deviation of LRM^{2}QAS, together with the mean absolute error between LRM^{2}QAS and human scores.

In this context, *Mean Error* denotes the average absolute difference between LRM^{2}QAS and human judgments within a given length bin and serves as a localised indicator of alignment. The results

Table 8: KC–human correlation conditioned on answer length.

Length Bin	Count	Mean KC	KC–Human Pearson
Short ($\leq$20)	504	0.27	0.52
Medium (20–50)	119	0.31	0.42
Long (>50)	17	0.49	0.57

show an increase in mean LRM^{2}QAS from short to medium-length answers, reflecting improved coverage of core factual content as answer length increases. For long answers, the mean LRM^{2}QAS score exhibits saturation while variance increases, indicating greater sensitivity to how additional information is organised and expressed. Elevated mean error values for medium and long answers suggest that content beyond essential factual units contributes more strongly to human judgments than to the components emphasised by LRM^{2}QAS.

Figure 6 illustrates this saturation trend, showing diminishing gains in LRM^{2}QAS as answer length increases.

Table 9: Length-conditioned stability analysis of LRM^{2}QAS.

Length Bin	Count	Mean LRM^{2}QAS	Std. LRM^{2}QAS	Mean Error
Short	215	0.16	0.18	56.35
Medium	215	0.31	0.24	65.51
Long	210	0.29	0.21	57.85

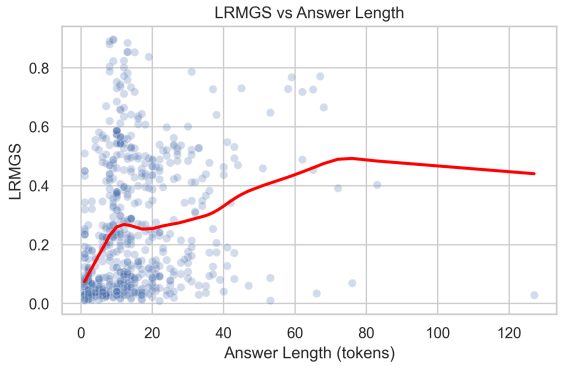

Figure 6: LRM^{2}QAS scores as a function of answer length, illustrating saturation effects for longer outputs.

G.4 Cross-Lingual Robustness and Translation Sensitivity

Cross-lingual robustness is evaluated using a language-by-system heatmap of mean LRM^{2}QAS scores, shown in Figure 7. Although absolute metric values vary across languages, relative system rankings remain largely consistent, indicating stability of nugget matching under translation-based

English pivoting. Table 10 further reports language-wise correlations between LRM^{2}QAS and human scores.

Lower correlation values for certain languages coincide with increased lexical divergence introduced by translation, suggesting sensitivity to paraphrastic variation rather than metric instability. No systematic degradation is observed for ultra-low-resource languages, indicating that LRM^{2}QAS retains discriminative capacity despite translation noise.

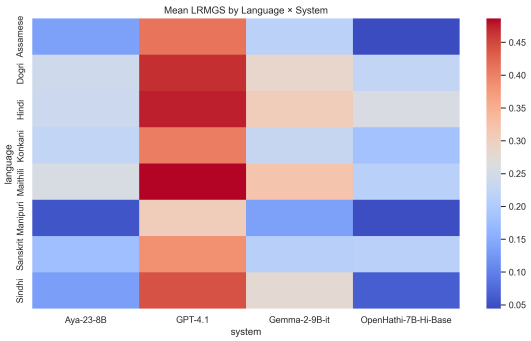

Figure 7: Mean LRM^{2}QAS scores across languages and systems.

Table 10: Language-level robustness analysis of LRM^{2}QAS.

Language	Count	Mean LRM^{2}QAS	Corr. w/ Human
Assamese	80	0.20	0.56
Dogri	80	0.30	0.49
Hindi	80	0.32	0.46
Konkani	80	0.26	0.38
Maithili	80	0.32	0.47
Manipuri	80	0.14	0.55
Sanskrit	80	0.25	0.40
Sindhi	80	0.23	0.59

Overall, the ablation results indicate that LRM^{2}QAS aligns most strongly with human judgments for short, factual answers, while predictable divergence emerges for longer or stylistically richer outputs. This behavior reflects the intended design of LRM^{2}QAS, which prioritizes factual nugget precision over discourse-level fluency or elaboration.