PARIKSHA: A Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data

Ishaan Watts Varun Gumma[♠] Aditya Yadavalli[♢]
Vivek Seshadri[♠♢] Manohar Swaminathan[♠] Sunayana Sitaram[♠]
[♠]Microsoft Corporation [♢]Karya

Elo Leaderboard

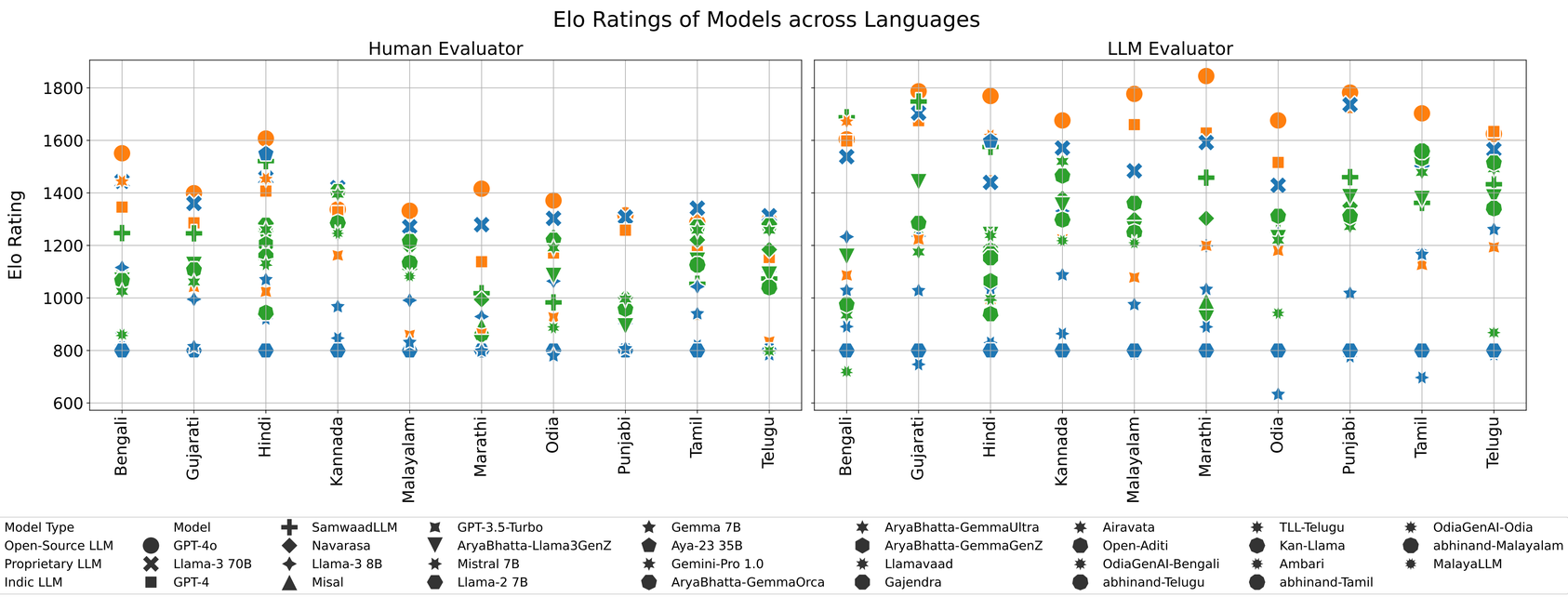

Figure 1. Comparison of Elo ratings of models across languages evaluated by both humans and an LLM. We group all models into three categories - Indic, Proprietary and Open-Source base LLMs.

Human-LLM Agreement

Prompt Type	Pairwise		Direct	
	$\mathcal{H}$-$\mathcal{H}$	$\mathcal{H}$-LLM	$\mathcal{H}$-$\mathcal{H}$	$\mathcal{H}$-LLM
All	0.54	0.49	0.49	0.31
Cultural	0.50	0.44	0.47	0.24
Non-Cultural	**0.57**	**0.55**	**0.49**	**0.37**

Table 1. Average Fleiss Kappa (κ) coefficient between Humans and Human-LLM. Here $\mathcal{H}$ stands for Humans.

Language	Pairwise	Direct
Average	0.76	0.65
Bengali	0.66	0.43
Gujarati	0.85	0.75
Hindi	0.80	0.67
Kannada	0.76	0.55
Malayalam	0.82	0.66
Marathi	0.82	0.82
Odia	0.78	0.53
Punjabi	0.69	0.54
Tamil	0.71	0.60
Telugu	0.70	0.91

Table 2. Kendall Tau (τ) correlations between Elo and DA leaderboards constructed by humans and LLM evaluator.

Motivation

- **Gap in Multilingual Performance of LLMs:** Underperformance in non-English, non-Latin script, and under-resourced languages.
- **Test Dataset Contamination:** Popular benchmarks already consumed in LLM training data, compromising fair evaluation.
- **Lack of Linguistic and Cultural Context in Translated Benchmarks:** Limited benchmarks, often translated, lacking essential linguistic and cultural nuances.
- **Difficulty in Scaling Human Evaluations:** Absence of subjective metrics; human evaluations are costly and time-intensive.

Contributions

- **Extensive Human Evaluation:** 90K evaluations across 10 Indic languages, comparing 30 Indic and multilingual models on a culturally-nuanced dataset.
- **LLM-Based Evaluator Analysis:** Comprehensive evaluation of human-LLM agreement in multilingual settings, the largest analysis of its kind.
- **Leaderboard Creation and Analysis:** Leaderboards based on human and LLM evaluations, highlighting trends and biases across languages and models.

Evaluation Pipeline

We evaluate the models on **Open-ended Question Answering** using **2 different strategies** (Pairwise-Assessment and Direct Assessment) and by **2 types of evaluators** (Human and LLM).

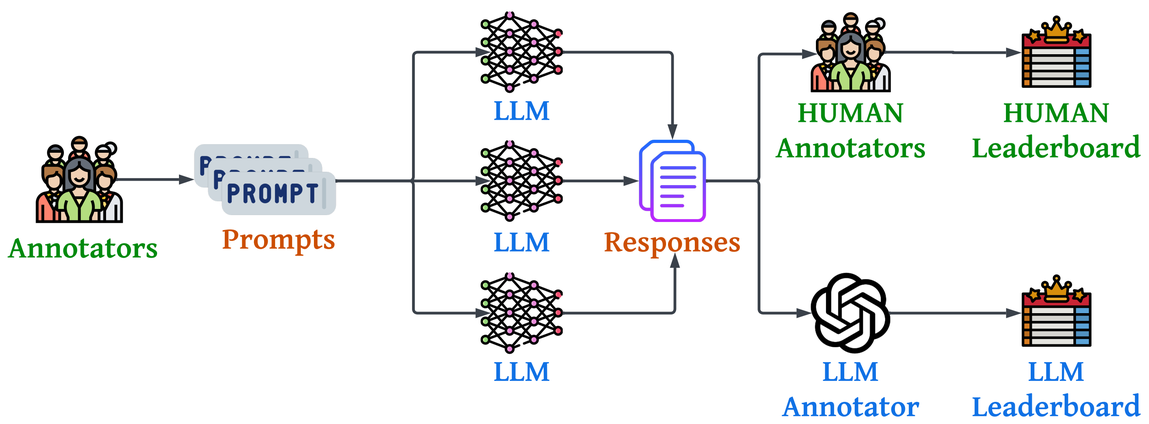

Figure 2. Evaluation pipeline

Key Takeaways

RQ1. Competitiveness of Indic LLMs with Proprietary Models:

- Smaller Indic models outperform their open-source bases, while large models like GPT-4o and Llama-3 70B excel in Indic languages.
- Newer open-source models, such as Llama-3, show strong potential for fine-tuning in Indic languages.

RQ2. Feasibility of LLM Evaluators as Human Substitutes in Multilingual Settings:

- LLM evaluators align more closely with humans in pairwise (Elo) tasks than in direct assessment (DA).
- They struggle with culturally nuanced responses, especially in languages like Bengali and Odia.
- LLM evaluators capture high-level trends effectively, as seen by the τ scores in Table 2.

RQ3. Biases Affecting Evaluator Judgments:

- **Position Bias:** No position bias observed in LLM evaluators through option flipping analysis.
- **Verbosity Bias:** Both human and LLM evaluators show a slight preference for longer responses.
- **Option Bias:** LLMs display optimism bias, with higher scores in DA and fewer ties in pairwise evaluations; struggle with hallucination detection.
- **Self Bias:** GPT-4 evaluators tend to prefer their own outputs.

Prompt Curation

We include the following **10 Indian languages** in our evaluation: **Hindi, Tamil, Telugu, Malayalam, Kannada, Marathi, Odia, Bengali, Gujarati, and Punjabi**.

Prompt Type	Examples
Finance (5)	*What is the difference between a debit card and a credit card?*
Health (5)	*How can I improve my posture to prevent back and neck pain?*
Cultural (10)	(Kannada) *Although in the neighboring states movie actors rise to prominence in politics, why is it not seen in Karnataka?* (Telugu) *In Telugu tradition, why do parents of girls pay for the first birth of a child?*

Table 3. Table containing number of prompts in each category per language with examples (English translation for readability).

Safety Evaluation

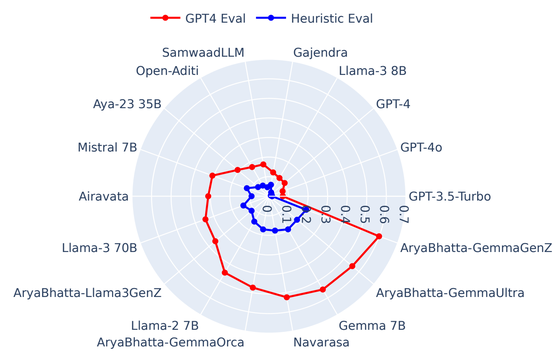

Figure 3. RTP-LX Safety Evaluation of Hindi models.

Evaluation Statistics

Language	Models	Pairwise	Direct
All	30 (20+10)	21690	8640
Hindi	20 (10+10)	4180	1200
Telugu	15 (7+8)	2310	900
Bengali	15 (6+9)	2310	900
Malayalam	14 (6+8)	2002	840
Kannada	14 (6+8)	2002	840
Tamil	14 (6+8)	2002	840
Odia	14 (6+8)	2002	840
Gujarati	13 (5+8)	1715	780
Punjabi	13 (5+8)	1715	780
Marathi	12 (4+8)	1452	720

Table 4. Number of pairwise comparison (battle) and direct assessment datapoints for each language.

Biases in Evaluation

Self Bias: *Average rank of GPT-4 increases by the highest amount (1.4 places) across Elo Leaderboards for evaluations performed by the GPT evaluator.*

Position Bias

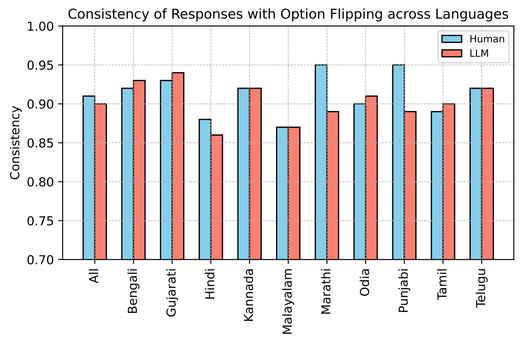

Figure 4. Consistency of response with option flipping across languages for humans and LLM evaluator.

Verbosity Bias

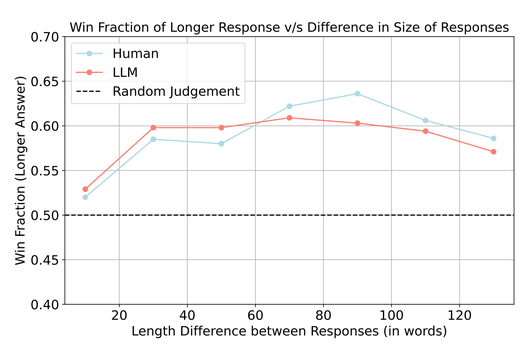

Figure 6. Figure showing the win fraction of a longer answer over a shorter answer.

Option Bias (Elo)

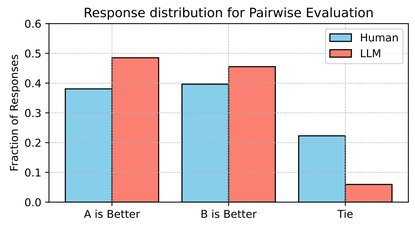

Figure 5. Response distribution for humans and LLM evaluator in Pairwise Evaluations.

Option Bias (DA)

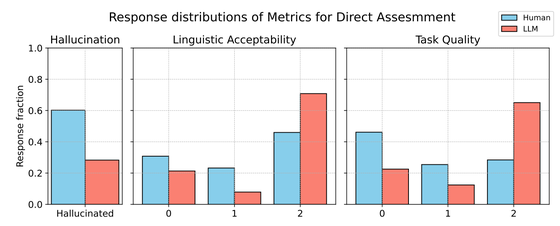

Figure 7. Response distribution of Hallucination, Linguistic Acceptability and Task Quality metrics for humans and LLM evaluator in DA.

wattsishaan18@gmail.com
EMNLP 2024, Miami
sunayana.sitaram@microsoft.com