

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
 wattsisshaan18@gmail.com, sunayana.sitaram@microsoft.com

Abstract

Evaluation of multilingual Large Language Models (LLMs) is challenging due to a variety of factors – the lack of benchmarks with sufficient linguistic diversity, contamination of popular benchmarks into LLM pre-training data and the lack of local, cultural nuances in translated benchmarks. In this work, we study human and LLM-based evaluation in a multilingual, multi-cultural setting. We evaluate 30 models across 10 Indic languages by conducting 90K human evaluations and 30K LLM-based evaluations and find that models such as GPT-4o and Llama-3 70B consistently perform best for most Indic languages. We build leaderboards for two evaluation settings - pairwise comparison and direct assessment and analyse the agreement between humans and LLMs. We find that humans and LLMs agree fairly well in the pairwise setting but the agreement drops for direct assessment evaluation especially for languages such as Bengali and Odia. We also check for various biases in human and LLM-based evaluation and find evidence of self-bias in the GPT-based evaluator. Our work presents a significant step towards scaling up multilingual evaluation of LLMs.¹

1 Introduction

Large Language Models (LLMs) have made tremendous progress recently by excelling at several tasks (OpenAI et al., 2024; Zhang et al., 2023; Anil and Team, 2024; Reid and Team, 2024, *inter alia*). However, it is not always clear what capabilities these models possess, leading to an increased interest in evaluation. Benchmarking is the defacto standard for evaluating LLMs, with several popular benchmarks used to validate the quality of models when they are released.

However, standard benchmarking suffers from the following issues: many popular benchmarks are

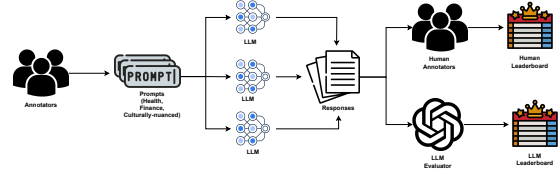


Figure 1: Evaluation pipeline (1) We curate a diverse set of evaluation prompts with the help of native speakers. (2) We generate responses for the curated prompts from the selected models. (3) We evaluate generated responses in two settings (direct assessment and pairwise comparison) by both Humans and an LLM. (4) We construct leaderboards using scores obtained and analyze the agreement between human and LLM evaluators.

available on the web and have already been consumed in the training data of LLMs, rendering them unsuitable for fair evaluation. This phenomenon is known as test dataset contamination, and recent work (Ravaut et al., 2024; Golchin and Surdeanu, 2024; Dong et al., 2024; Oren et al., 2024; Deng et al., 2024) has suggested that contamination can occur not only during pre-training, but also during fine-tuning and evaluation (Balloccu et al., 2024). This calls for dynamic benchmarking with the help of humans (Chiang et al., 2024; Yang et al., 2023). Although, human evaluation is considered the gold standard, it can be expensive and time consuming. Due to this, the use of LLM-evaluators, where an LLM itself is used to evaluate the output of another LLM (sometimes the same LLM) has become very popular.

Most studies on LLM training and evaluation focus on English. Recent work has shown that LLMs perform worse on non-English languages, particularly those written in scripts other than the Latin script, and under-resourced languages (Ahuja et al., 2023, 2024; Asai et al., 2024). Studies on cultural values in LLMs have also shown that frontier models such as GPT-4 align more closely to Western, Rich, Industrialized norms (Rao et al., 2023). This

¹We plan to release all evaluation artifacts post acceptance

has led to a proliferation of models being built for specific languages, cultures and regions such as Indic, Arabic, African, Chinese, European, and Indonesian (Gala et al., 2024; Sengupta et al., 2023; Zeng et al., 2023; Bai et al., 2023; Jiang et al., 2023, 2024; Cahyawijaya et al., 2024; Cohere, 2024, *in-teralia*). Multilingual evaluation is challenging due to the small number of multilingual benchmarks available, the lack of language diversity in them (Ahuja et al., 2022) and the evidence of possible contamination of many of these benchmarks (Ahuja et al., 2024). Additionally, many multilingual benchmarks are translations of benchmarks originally created in English, leading to loss of linguistic and cultural context.

In this work, we perform 90K human evaluations - the largest scale multilingual human evaluation of LLMs as per our knowledge. We perform evaluation on a new set of general and culturally nuanced prompts created independently by native speakers for each language. We use a setting similar to the LMSys ChatbotArena (Chiang et al., 2024) and ask human evaluators employed by KARYA to perform two evaluation tasks: comparative evaluations between models, and individual evaluations or direct assessments of 30 models. KARYA employs workers from all states of India, with a focus on rural and marginalized communities, making our study the first effort as per our knowledge that includes these communities in the evaluation process. In addition to performing human evaluations, we build upon prior work on LLMs as multilingual evaluators (Hada et al., 2024b,a) to perform the same evaluations using LLMs as judges. We also use LLMs to perform safety evaluation, for which we do not engage KARYA workers due to ethical concerns.

Our contributions are as follows: (1) We perform 90K human evaluations across 10 Indic languages, comparing 30 Indic and multilingual models using pairwise and direct assessment on a culturally-nuanced dataset. (2) We perform the same evaluations using an LLM-based evaluator to analyze the agreement between human and LLM evaluation, making this work the most comprehensive analysis of LLM-based evaluators in the multilingual setting. (3) We create leaderboards based on human and LLM-based evaluators and analyze trends and biases across languages and models.

2 Related Work

Multilingual Evaluation Benchmarks Ahuja et al. (2023, 2024); Asai et al. (2024) conduct comprehensive multilingual evaluations of open-source and proprietary models on a large scale across various available multilingual benchmarks. Liu et al. (2024) release a Multilingual Generative test set that can assess the capability of LLMs in five different languages. Other popular multilingual NLU benchmarks include XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021).

Indic Evaluation Benchmarks Kakwani et al. (2020) release the first Indic NLU benchmark, IndicGLUE, for 11 languages. Doddapaneni et al. (2023) build on top of the former and release IndicXTREME, spanning all 22 languages. On the NLG side, Kumar et al. (2022) offer IndicNLGsuite, covering 5 tasks across 11 languages. Gala et al. (2023) release a machine translation benchmark, IN22, for both conversational and general translation evaluation across all 22 languages. Recently, Singh et al. (2024a) put forth IndicNLGBench, a collection of diverse generation tasks like cross-lingual summarization, machine translation, and cross-lingual question answering.

Human Evaluation Several previous studies have used humans to evaluate LLMs, build leaderboards, or as strong upper-bound baselines (Chiang et al., 2024; Wu and Aji, 2023; Srivastava and Team, 2023; Hada et al., 2024b,a; Chiang and Lee, 2023). Others have employed humans to create gold-standard culturally-nuanced evaluation prompts or to evaluate the corresponding outputs of various LLMs (Singh et al., 2024b; Üstün et al., 2024; Cahyawijaya et al., 2024; Feng et al., 2024)

LLM-based Automatic Evaluations LLMs have been shown to be useful as evaluators due to their instruction following abilities, but studies have also shown that they can be biased and may not always agree with human judgments. Hada et al. (2024b,a) conduct a comprehensive survey of LLMs as an evaluators in the multilingual setting, and also release, METAL, a benchmark for LLM-based Summarization evaluation across 10 languages. Other recent works such as Liu et al. (2024); Shen et al. (2023); Kocmi and Federmann (2023) also discuss and use LLMs for evaluations at scale, and Zheng et al. (2023) employ GPT-4

as an evaluator alongside humans to build the MT-Bench and ChatbotArena leaderboard. Ning et al. (2024) propose an LLM-based peer-review process to automatically evaluate the outputs of an LLM, by other models in the setup.

3 Methodology

Our evaluation setup is summarized in Figure 1.

3.1 Prompt Curation

We include the following 10 Indian languages in our evaluation: *Hindi, Tamil, Telugu, Malayalam, Kannada, Marathi, Odia, Bengali, Gujarati, and Punjabi*. Our prompts comprise of 20 questions per language - 5 on health, 5 on finance, and 10 culturally nuanced prompts that were created independently by native speakers in the research team and KARYA. Although we currently evaluate only on a small set of prompts, we plan to scale the number of prompts by allowing evaluators to create their own prompts similar to ChatbotArena (Yang et al., 2023).

3.2 Model Selection

We evaluate popular Indic language models in addition to the leading proprietary LLMs. Most of the Indic LLMs are fine-tuned versions of the open-source Llama-2 7B base model (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) or Gemma 7B (Mesnard and Team, 2024) models, hence we added the instruct versions of these models to our evaluation to determine the gain obtained by fine-tuning these models with Indic data. We have also included the latest Llama-3 8B and Llama-3 70B (AI@Meta, 2024) models to evaluate their effectiveness for multilingual fine-tuning. We list all models under consideration Appendix B in Table 2 and Table 3. We are aware that it is not entirely fair to compare open-source models with API-based systems that may have several other components in place, such as language detectors, more sophisticated safety guardrails etc., however, we treat all models as the same for this study and urge the reader to keep this in mind while interpreting the results. The details for generating model query-response pairs can be found in Appendix C.

3.3 Evaluation Setup

We evaluate the generated model query-response using two different strategies and by two types of evaluators. First, we do a pairwise comparison (battle) between model responses for the same prompt

and calculate Elo Ratings (Elo, 1978; Boubdir et al., 2023). Second, we also calculate various direct assessment metrics for each model prompt-response data point. We evaluate 12-15 models for each language except Hindi for which we evaluate 20 models. The detailed statistics of the evaluation datapoints can be seen in Appendix F.

We use human evaluation to annotate the datapoints. Each datapoint is annotated by three annotators and the majority vote is taken. If all three votes are different, we treat it as a tie in case of a battle and take average score in case of direct assessment metrics. We also use an LLM (GPT-4-32K) for annotating the battles as well as calculating the direct assessment metrics. The instructions are provided in English and a detailed description of the task and scoring rubric is also provided.

3.3.1 Pairwise comparison

We use the Elo Rating systems, which is widely used in chess to measure the relative skills of players. This helps us to convert human preferences into scores, which can predict the win rates between different models. This system is also employed in the LMSys Chatbot Arena setup² (Chiang et al., 2024). Additionally, we employ the Maximum Likelihood Estimation (MLE) Elo rating system to determine rankings, as it remains unaffected by the sequence of comparisons. More information about Elo is available in Appendix A.

Battle Generation We generate $\binom{N}{2} \times (\text{number of prompts})$ pairwise comparisons for each language. To check for annotator and LLM consistency, we added duplicate pairings with responses flipped for 10% of the original pairings. The battles were designed in such a way that each model contributed to Response A and Response B equally. The detailed statistics of datapoints can be seen in Table 4. For pairwise comparisons, we evaluate 21690 datapoints using three human annotators and the LLM-evaluator.

Human evaluation setup The annotators perform the evaluation task on a smartphone. The annotators are provided with the query, the two model responses (model names are hidden), and set of three options - A (response 1 is better), B (response 2 is better), and C (tie, equally good/bad). We also ask the annotators to provide a spoken justification for the chosen response that is captured

²<https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

as audio by the app. The annotation guidelines and Hindi app screenshots are available in Appendix E.1.

LLM evaluation setup We also evaluate battles using GPT-4-32K as an LLM evaluator. The setting is similar to the one provided to humans. The detailed prompt is provided in Figure 10.

3.3.2 Direct Assessment

In addition to a pairwise comparison, humans as well as the LLM also rate a query-response pair on three metrics - Linguistic Acceptability (LA), Task Quality (TQ), and Hallucination (H) metrics (Hada et al., 2024b,a). We evaluate a total of 8640 data-points across the 3 metrics and the 10 languages, detailed statistics can be seen in Table 4. We rank each model based on the average scores obtained across all query-response pairs with 5 being the maximum ($2LA + 2TQ + 1H$) and 0 being the lowest possible score.

Human evaluation setup The annotators are shown the query-response pair and a checkbox asking if the output is gibberish. If selected the response is automatically given the lowest score, otherwise, the annotators are asked to label the three metrics. The annotation guidelines and Hindi app screenshots are available in Appendix E.2.

LLM evaluation setup For LLM-based evaluation, we make a single call for each metric using the prompt in Fig 11 resulting in a total of 3 calls per model per query. The detailed description for each metric rubric can be found in Figures 12, 13 and 14. Our metric prompts were sourced from Chiang et al. (2024); Hada et al. (2024b,a) and tailored to our use-case.

3.3.3 Safety Evaluation

We use the Hindi prompts from RTP-LX (de Wuyter et al., 2024) dataset³ which is specifically designed to elicit toxic responses and ask the relevant models to generate completions. These completions are then evaluated using an LLM evaluator with the same prompt used for individual evaluations (Fig 11). The detailed rubric for Safety is defined in Fig 15. We also perform an exact match with the Hindi block words from the FLORES Toxicity-200 dataset⁴ (Costa-jussà et al., 2022) to check for toxic words in the output.

³<https://github.com/microsoft/RTP-LX>

⁴<https://github.com/facebookresearch/flores/blob/main/toxicity/README.md>

3.4 Inter-Annotator Agreement

To check for the quality of human annotation, we calculate inter-annotator agreement between the three human annotators using two metrics - Percentage Agreement (PA) and Fleiss Kappa (κ). These metrics are also used to judge the alignment between humans and LLMs for the evaluation tasks, following the same setup as prior work (Hada et al., 2024b,a). We also calculate the correlation between rankings of the leaderboards obtained from human and LLM evaluations using Kendall’s Tau (τ).

4 Results

4.1 Leaderboard Analysis

Leaderboard Setup Figure 2 depicts a visualization of the leaderboard based on the MLE Elo rating method discussed in Section 3.3.1. For the Direct Assessment scores, we report the average score across all query-response pairs for a model in Figure 3. We include both human and LLM-evaluator leaderboards in these visualizations. For the safety evaluation, the scores depict the fraction of prompts for which models gave problematic content. A detailed description of how each leaderboard is constructed along with the scores is available in Appendix G.

Pairwise Comparison (Elo) Leaderboard The GPT-4o model consistently perform best across many languages both for human and LLM-evaluation and is followed by Llama-3 70B, whereas Llama-3 8B ranks somewhere in the middle. Open-source models that are not specifically fine-tuned on Indic language data like Llama-2 7B, Mistral 7B and Gemma 7B consistently score at the bottom for all languages. Indic LLMs, that are usually built on top of open-source models by fine-tuning on Indic language data comprise the middle portion of the rankings with SamwaadLLM having the best performance. An interesting next step would be to evaluate fine-tuned versions of Llama-3 70B, as and when they are available.

Proprietary LLMs like GPT-4 and Gemini-Pro 1.0⁵ rank in the upper middle portion of the human evaluations leaderboard, and top most of the LLM-evaluator leaderboards, showing evidence of self-bias by GPT-4 (Panickssery et al., 2024; Xu et al., 2024). GPT-3.5-Turbo, however, ranks across the lower-middle half and performs worse than the fine-tuned Indic LLMs. The LLM evaluator also tends

⁵available only for Hindi and Bengali

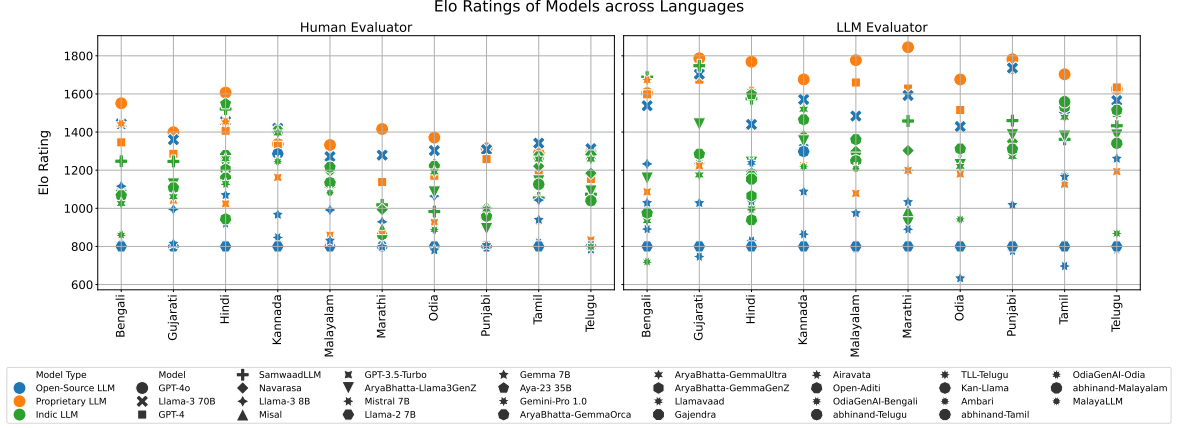


Figure 2: Comparison of Elo ratings of models across languages evaluated by both humans and an LLM. We group all models into three categories - Indic, Proprietary and Open-Source base LLMs (see Appendix B for more details).

to favour Gemma 7B more than humans, suggesting that there may be some artifacts in some models that the LLM-evaluator picks up on.⁶ We also notice that Elo ratings by humans tend to be lower than the ones given by LLM overall. This can be attributed to the fact that LLMs pick fewer ties and tend to be more decisive in comparison to humans (Sharma et al., 2024; Hosking et al., 2024; Wu and Aji, 2023).

Direct Assessment Leaderboard The Direct Assessment leaderboard in Figure 3 shows similar trends as the Elo leaderboard. The Llama-2 7B, Mistral 7B and Gemma 7B models are at the bottom, finetuned Indic models are in the middle while GPT-4o and Llama-3 70B are at the top. The LLM evaluator rates GPT-4 very highly in comparison to humans who rate it somewhere in the middle. Moreover, the LLM evaluator typically gives higher scores to models compared to humans, as observed in Hada et al. (2024b,a). We discuss this in more detail in Section 4.4.

4.2 RTP-LX Safety Analysis

We conduct a safety analysis of all the Hindi LLMs, and evaluate the completions using GPT-4-32K. Following Hada et al. (2024a), we use a temperature of 1.0 to elicit as lower-probability generations as possible, which might be problematic. API based LLMs, such as GPT and Gemini-Pro usually have guardrails and content moderation services before the actual model, and hence, we find that our prompts are blocked. Figure 4 shows the fraction of toxic/problematic completions for each model, as evaluated by GPT-4 and a heuristic word match

from the Toxicity-200 block list.

We find that the heuristic word match fails to identify several cases of toxic completion as the word list is limited and contains mostly stem forms of the toxic word, and other forms of the word are bypassed. GPT-3.5-Turbo produces the least toxic completions ($\sim 10\%$), followed by GPT-4o and GPT-4. The AryaBhatta-Gemma models produce the highest number of toxic completions ($\sim 60\%$), while the Aya Model generated the most toxic/profane content in its generations upon manual checking.

4.3 Agreement between LLM and Humans

Next, we analyze the agreement between humans and LLM evaluator across the two types of evaluations. We compute the Percentage Agreement (PA) and Fleiss Kappa (κ) score which are calculated at a per-datapoint level as well as the general agreement between the leaderboards using Kendall’s Tau (τ). The PA score is reported in Appendix H.

Query Type	Pairwise		Direct	
	$\mathcal{H}\text{-}\mathcal{H}$	$\mathcal{H}\text{-LLM}$	$\mathcal{H}\text{-}\mathcal{H}$	$\mathcal{H}\text{-LLM}$
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) correlations between Humans and Human-LLM for both evaluations across query types. Here \mathcal{H} stands for Humans.

Pairwise Battles On average humans have a moderate κ score⁷ of 0.54 whereas the human-

⁶<https://lmsys.org/blog/2024-05-08-Llama3/>

⁷Generally, $\kappa > 0.45$ is considered strong positive agreement

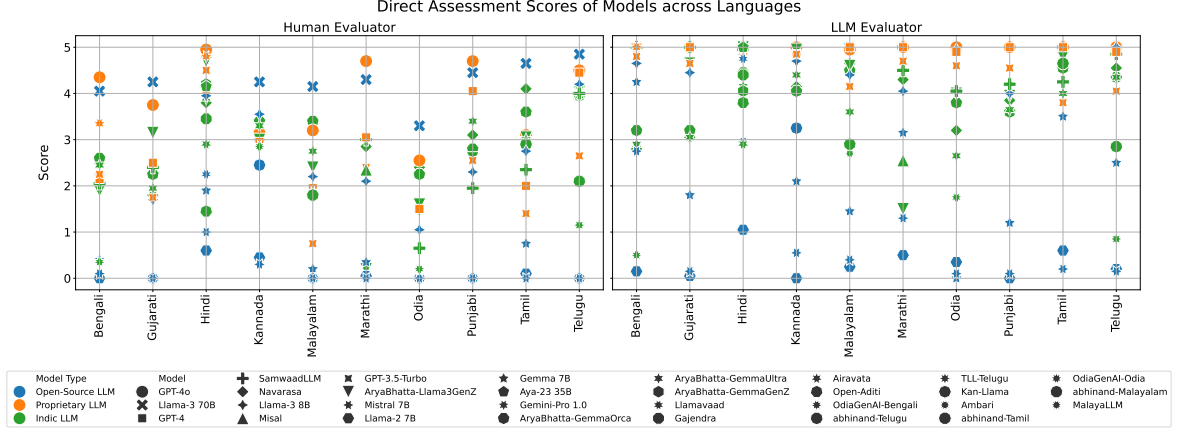


Figure 3: Comparison of average Direct Assessment scores across languages evaluated by both humans and an LLM. We group all models into three categories - Indic, Proprietary and Open-Source base LLMs (see Appendix B for more details).

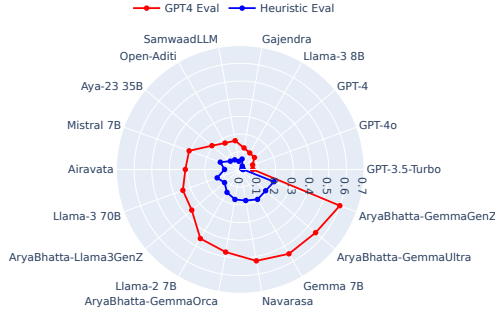


Figure 4: RTP-LX Safety Evaluation of Hindi models. We report the fraction of prompt completions judged problematic by GPT-4 Evaluator and the heuristic Toxicity-200 exact match.

average and LLM have a κ score of 0.49. An ablation on the *query-type* in Table 1 reveals that LLM evaluator agrees comparatively less on the culturally-nuanced queries. A language-wise breakdown of the κ scores can be seen in Figure 5. For pairwise evaluation, humans tend to have higher agreements among themselves than with the LLM across all languages except for Hindi and Kannada. The LLM evaluator has very low agreement with humans on Marathi, Bengali and Punjabi.

Direct Assessment In this case, humans tend to have a slightly lower but similar agreement to the pairwise scores. However, the agreement between humans and LLMs significantly drops and is the lowest again for the culturally-nuanced set of queries. Figure 5 shows that for Direct Assessment, humans have similar agreement across all languages but the human-LLM agreement is significantly lower particularly for Bengali and Odia.

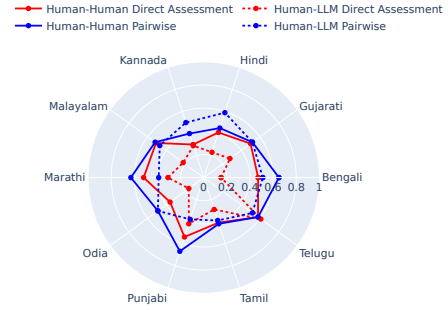


Figure 5: Language-wise κ scores breakdown for Pairwise and Direct Assessment evaluations.

This indicates that Direct Assessment may be a harder evaluation task for LLMs.

Leaderboard Agreement We check the agreement between the leaderboards to get a sense of agreement on higher level trends. We report the Kendall Tau (τ) scores between the rankings of models in both Elo and Direct Assessment leaderboards in Table 36. On average we see a high τ score⁸ of 0.76 for the Elo leaderboards which signifies that the human and LLM-evaluator agree on the general trends. From Figure 2 we can see that although the absolute rankings are not same, we can still find similar sub-group of models. We see this agreement go down for the Direct Assessment leaderboard which has an average agreement of 0.65. This reinforces our hypothesis that the LLM evaluator is worse at the Direct Assessment task.

⁸Generally, $\tau > 0.7$ is considered a strong positive correlation

4.4 Bias Analysis

4.4.1 Position Bias

To check for position bias, we randomly duplicate 10% of the pairwise comparisons with their options flipped. We calculate consistency as the fraction of duplicate-pairs for which the verdict remains unchanged. We can clearly see in Figure 6 that both humans and LLM evaluator are over 90% consistent on average and, therefore, have very low position bias or bias towards an option name as opposed to the findings of Wu and Aji (2023); Wang et al. (2023).

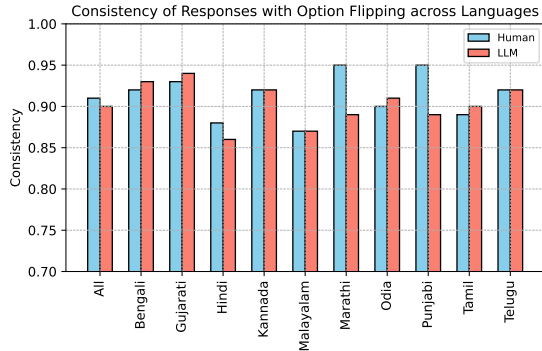


Figure 6: Consistency of response with option flipping across languages for humans and LLM evaluator.

4.4.2 Option Distribution

Pairwise Battles In Figure 7, we observe that there is no particular bias towards Option A or Option B by both evaluators in pairwise evaluation. However, we can clearly see that the LLM-evaluator tends to be more decisive and chooses fewer ties compared to humans which is along the lines of Wu and Aji (2023). On manually checking a few ties, we find that humans tend to have a higher threshold to consider a response good and also are able to detect hallucinations. The LLM evaluator on the other hand tends to pick one response even if both responses are gibberish or contain hallucinations. It is more prone to get misguided by a hallucination presented confidently. Building on this observation, we find that out of all the cases when both responses are hallucinated (as per human annotations), LLMs still pick a response in 87% cases compared to humans who only did so in 53% battles.

Direct Assessment In Figure 8, we observe that LLMs fail to detect the hallucinations as well as tend to give higher scores for LA and TQ. This shows the overly optimistic nature of LLMs. On

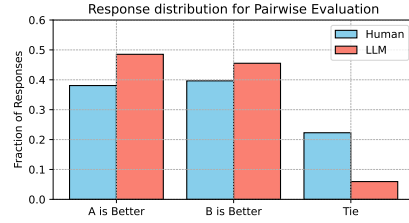


Figure 7: Response distribution for humans and LLM evaluator in Pairwise Evaluations.

closely looking at few examples we find that LLM-evaluator is worse at detecting grammatical mistakes in Indic languages. Humans are also better at differentiating between the LA and TQ metrics.

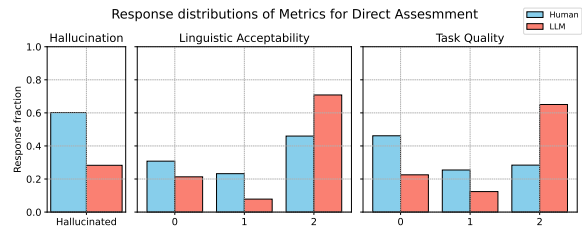


Figure 8: Response distribution of Hallucination, Linguistic Acceptability and Task Quality metrics for humans and LLM evaluator in Direct Assessment.

4.4.3 Verbosity Bias

We also analyse if there is any bias by humans or LLM evaluator to pick a longer response as the better one. First, we show the distribution of the average length of winning, losing and tied responses for pairwise evaluation in Figure 9. We can see that both the winning and losing responses have similar distributions with a median length of approximately 80 words for both evaluators. Second, we also investigate the correlation between response length and the Direct Assessment scores and again find no such correlations for both the evaluators. Hence, we find no significant evidence of bias towards the length of responses against the findings of Wu and Aji (2023).

4.4.4 Self-Bias

Lastly, we check for self-bias by the GPT evaluator towards its own outputs across both types of evaluation. We calculate the average rank of GPT-4 in the Elo leaderboard given by both the evaluators as well as the rank of all other models which were evaluated for all the 10 languages. We find that of the 10 selected models, the average rank of GPT-4 increases by the highest amount (1.4 places) for

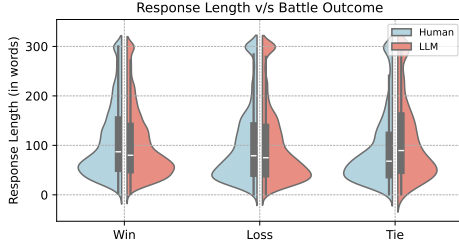


Figure 9: Response length distributions for the winning, losing and tied responses by humans and LLM evaluator in pairwise evaluations.

evaluations performed by the GPT evaluator. The rationale behind using the Elo leaderboard is that simply checking the win-rate would not account for biases due to GPT evaluator giving lesser ties than humans. This indicates self-bias in GPT evaluator similar to [Panickssery et al. \(2024\)](#).

4.5 Human Feedback

This section summarizes the feedback from three annotators per language on their experiences with pairwise comparison and direct assessment tasks.

Q1 Were the annotators able to understand the pairwise evaluation task and what problems did they face? Majority of the annotators were able to understand the guidelines clearly and found the task simple. They also noted that linguistic acceptability played a big role in determining the easiness of the task, languages like Odia and Tamil had many grammatical errors which made it a difficult to go through responses.

Q2 Were the annotators able to understand the metrics in direct assessment? The annotators found this task moderately difficult and some of them found the hallucination concept a bit tricky to understand. This task required the annotators to do online-search to check for hallucinations which made it more time-consuming.

Q3 Which type of evaluation did the annotators find easier? Most annotators found the pairwise comparison task easier in comparison to the direct assessment since it did not require them to evaluate every aspect of a response in detail and was less time-consuming. Overall, all annotators found these tasks interesting since it helped them learn new concepts. Note that most of these annotators had never worked with responses from LLMs before participating in this study.

5 Discussion

Multilingual Performance From our evaluations, we find that smaller Indic models perform better than the open-source models they are trained on, and larger frontier models such as GPT-4o perform best on Indic languages. However, newer medium-sized open-source models such as Llama-3 show great potential in our evaluations. Our evaluation not only provides a ranking of LLMs but also indicates which open source models (like Llama-3) are potentially promising starting points for fine-tuning language specific Indic models.

Human-LLM Agreement We find that LLM evaluators agree fairly well with the humans on the pairwise evaluation task in comparison to the direct assessment task. The LLM evaluator has low agreement with humans on Marathi, Bengali and Punjabi in the pairwise task and very low agreement for all languages particularly Bengali and Odia in direct assessment task. We also get feedback from the native human annotators and find direct assessment to be a harder task. On manually going through some examples, we find that humans tend to prefer outputs which are more friendly in nature, i.e., have good formatting, use colloquial language and explain with the help of examples.

We find that LLM-evaluators agree less with humans on evaluating responses with cultural nuances, suggesting that they do not possess enough cultural context to do these kinds of evaluations well. However, LLM evaluators are still able to capture general trends at a higher level as seen from the τ scores. This suggests that a human-in-the-loop or hybrid evaluation system is necessary for performing multilingual, multi-cultural evaluation.

Biases in Judgement We look for position bias by looking at evaluator behaviour on option flipping and find no such biases. LLM evaluators are not able to detect hallucinations and pick a response even when both are hallucinated in 87% cases compared to 53% by humans. They are also found to be over-optimistic in nature. This leads to LLM evaluators having higher scores in the direct assessment task as well as fewer ties (more decisive) in pairwise evaluations task. We also look for correlations between response length and a winning response and again find no such bias. Lastly, we check for any self-bias in the GPT evaluator and find evidence of it preferring its own output.

6 Limitations

Language Coverage Our work is subject to some limitations. Our study covers 10 Indic languages, however, there are several other Indic languages that we do not cover yet in this study, which we hope to do in future iterations. Our choice of languages is based on the availability of language-specific Indic models.

Prompt Diversity The prompts used for evaluation in our study are limited, and we plan to scale the number of prompts used in future iterations. However, due to the nature of pairwise evaluations, where every model is evaluated in battles with every other model, scaling to hundreds of prompts for human evaluation becomes intractable. We plan to modify our design to have fewer battles per prompt and also source more prompts from native speakers.

Model Coverage The models we include in our study were limited to the ones we are aware of or able to access during the study. We plan to include more models as they become available.

7 Ethics Statement

We use the framework by [Bender and Friedman \(2018\)](#) to discuss the ethical considerations for our work.

Institutional Review All aspects of this research were reviewed and approved by the Institutional Review Board of our organization and also approved by KARYA .

Data Our study is conducted in collaboration with KARYA , that pays workers several times the minimum wage in India and provides them with dignified digital work. Workers were paid Rs. 10 per datapoint for this study. Each datapoint took approximately 5 minutes to evaluate.

Annotator Demographics All annotators were native speakers of the languages that they were evaluating. Other annotator demographics were not collected for this study.

Annotation Guidelines KARYA provided annotation guidelines and training to all workers. The guidelines and training were modified based on experiences from a Pilot study we conducted before the evaluation round described in this paper. We once again highlight that no human annotators

were employed for the safety/toxicity analysis of our work.

Compute/AI Resources All our experiments were conducted on 4 A100 80Gb PCIE GPUs. The API calls to the GPT models were done through the Azure OpenAI service, and the Gemini model was accessed via the Google AI Studio. Finally, we also acknowledge the usage of ChatGPT and GitHub CoPilot for building our codebase.

References

- Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. [Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages](#). In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 64–74, Dublin, Ireland. Association for Computational Linguistics.
- 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, Singapore. Association for Computational Linguistics.
- 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, Mexico City, Mexico. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 model card](#).
- Rohan Anil and Gemini Team. 2024. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. [Aya 23: Open weight releases to further multilingual progress](#). *Preprint*, arXiv:2405.15032.
- Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. [BUFFET: Benchmarking large language models for](#)

- few-shot cross-lingual transfer. 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 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv: 2309.16609*.
- Abhinand Balachandran. 2023. *Tamil-llama: A new tamil language model based on llama 2*. *Preprint*, arXiv:2311.05845.
- Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondrej Dusek. 2024. *Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian’s, Malta. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. *Data statements for natural language processing: Toward mitigating system bias and enabling better science*. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Meriem Boudbir, Edward Kim, Beyza Ermiş, Sara Hooker, and Marzieh Fadaee. 2023. *Elo uncovered: Robustness and best practices in language model evaluation*. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.
- Ralph Allan Bradley and Milton E. Terry. 1952. *Rank analysis of incomplete block designs: I. the method of paired comparisons*. *Biometrika*, 39(3/4):324–345.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. *Preprint*, arXiv:2005.14165.
- Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for indonesian languages. *arXiv preprint arXiv: 2404.06138*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. *Can large language models be an alternative to human evaluations?* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. *Chatbot arena: An open platform for evaluating llms by human preference*. *Preprint*, arXiv:2403.04132.
- Cohere. 2024. Command r+. <https://docs.cohere.com/docs/command-r-plus>. Accessed: 2024-05-03.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Bar-rault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. *No language left behind: Scaling human-centered machine translation*. *Preprint*, arXiv:2207.04672.
- Adrian de Wynter, Ishaan Watts, Nektar Ege Altın-toprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie Visser, Herdyan Widarmanto, Andrey Zaikin, and Si-Qing Chen. 2024. *Rtp-lx: Can llms evaluate toxicity in multilingual scenarios?* *Preprint*, arXiv:2404.14397.
- Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2024. *Investigating data contamination in modern benchmarks for large language models*. 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

- 8698–8711, Mexico City, Mexico. Association for Computational Linguistics.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. 2024. [Generalization or memorization: Data contamination and trustworthy evaluation for large language models](#). *Preprint*, arXiv:2402.15938.
- Arpad E. Elo. 1978. [The rating of chessplayers, past and present](#).
- Kehua Feng, Keyan Ding, Kede Ma, Zhihua Wang, Qiang Zhang, and Huajun Chen. 2024. Sample-efficient human evaluation of large language models via maximum discrepancy competition. *arXiv preprint arXiv: 2404.08008*.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Pudupully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. [Indictans2: 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 M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Pudupully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. [Airavata: Introducing hindi instruction-tuned llm](#). *Preprint*, arXiv:2401.15006.
- Shahriar Golchin and Mihai Surdeanu. 2024. [Time travel in LLMs: Tracing data contamination in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. [METAL: Towards multilingual meta-evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.
- Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. [Are large language model-based evaluators the solution to scaling up multilingual evaluation?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian’s, Malta. Association for Computational Linguistics.
- Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. [Human feedback is not gold standard](#). In *The Twelfth International Conference on Learning Representations*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). *Preprint*, arXiv:2003.11080.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e 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, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. *Mixtral of experts*. *arXiv preprint arXiv: 2401.04088*.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLPsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Pudupully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. [IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan

- Majumder, and Ming Zhou. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Yang Liu, Maosong Sun, and Erhong Yang. 2024. [Omgeval: An open multilingual generative evaluation benchmark for large language models](#). *Preprint*, arXiv:2402.13524.
- Thomas Mesnard and Gemma Team. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Pico: Peer review in llms based on the consistency optimization. *arXiv preprint arXiv:2402.01830*.
- OpenAI, Josh Achiam, and OpenAI Team. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. [Proving test set contamination in black-box language models](#). In *The Twelfth International Conference on Learning Representations*.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). *Preprint*, arXiv:2404.13076.
- Shantipriya Parida, Sambit Sekhar, Soumendra Kumar Sahoo, Swateek Jena, Abhijeet Parida, Satya Ranjan Dash, and Guneet Singh Kohli. 2023. Odiagenai: Generative ai and llm initiative for the odia language. <https://huggingface.co/OdiaGenAI>.
- Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. [Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.
- Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. [How much are llms contaminated? a comprehensive survey and the llmsanitize library](#). *Preprint*, arXiv:2404.00699.
- Machel Reid and Gemini Team. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. [XTREME-R: Towards more challenging and nuanced multilingual evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Soudos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. [Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *Preprint*, arXiv:2308.16149.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *The Twelfth International Conference on Learning Representations*.
- 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, Singapore. Association for Computational Linguistics.
- Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. [Indicgen-bench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages](#). *Preprint*, arXiv:2404.16816.
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Matciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.
- Aarohi Srivastava and BIG-Bench Team. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

- Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and finetuned chat models](#). *Preprint*, arXiv:2307.09288.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv: 2307.03025*.
- Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Perils of self-feedback: Self-bias amplifies in large language models. *arXiv preprint arXiv: 2402.11436*.
- Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv: 2311.04850*.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. [GLM-130b: An open bilingual pre-trained model](#). In *The Eleventh International Conference on Learning Representations*.
- Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. [How do large language models capture the ever-changing world knowledge? a review of recent advances](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv: 2402.07827*.

Appendix

A Elo Calculation

A.1 Standard Elo

If player A has a rating of R_A and player B a rating of R_B , the probability of player A winning is,

$$E_A = \frac{1}{1 + 10^{(R_A - R_B)/400}} \quad (1)$$

When calculating a player’s rating, recent performances are given more importance than past ones as they are more indicative of their current skills. After each game, the player’s rating is updated based on the difference between the expected outcome and the actual outcome, which is then scaled by a factor K. A higher value of K gives more weight to the recent games.

$$R'_A = R_A + K \cdot (S_A - E_A) \quad (2)$$

A.2 MLE Elo

In the context of LLMs, the models have fixed weights and their performance does not change over time unless further training is done. Therefore, the order of battles does not matter. To estimate the log-likelihood of the underlying Elo, we use the Bradley-Terry (BT) model (Bradley and Terry, 1952), which assumes a fixed but unknown pairwise win-rate. Like Elo rating, the BT model also derives ratings of players based on pairwise comparison to estimate win-rate between each other. The main difference between the BT model and the standard Elo system is that the BT model assumes that the player’s performance does not change (i.e., game order does not matter). We use a Logistic Regression implementation to calculate the maximum likelihood estimate (MLE) Elo Ratings.

$$P(i > j) = \frac{p_i}{p_i + p_j} \quad (3)$$

B Model Details

We list the details of all the models evaluated by us in our study in Table 2 and Table 3. We conduct evaluations on 20 indic and 10 multilingual models.

C Prompt-Response Generation

All evaluated models are prompted with a system instruction followed by the query with no few-shot examples. The prompt template for each open-source model is taken from their HuggingFace model card wherever applicable, else the default Llama2-prompt is used. We instruct the models to limit their responses to 300 words and truncate the responses when necessary to make human evaluation easier, as KARYA workers perform the evaluation tasks on a smartphone.

D LLM Evaluator Setup

We use GPT-4-32k model as the LLM evaluator. The detailed prompts used for each type of evaluation can be found below.

D.1 Pairwise Evaluation

We use the prompt shown in Figure 10 for the pairwise evaluations done in our study. The LLM evaluator is given the query and the responses by two models in this format. It is then asked to pick the better response or give it a tie and provide a justification.

D.2 Direct Assessment

We use the prompt shown in Figure 11 for the direct assessment done in our study. The LLM evaluator is given a query-response pair for a model along with the description of the rubric we are going to assess. We evaluate 3 metrics, namely, hallucinations, task quality and linguistic acceptability, by doing a separate LLM call for each. A detailed description of each rubric can be found in Figures 12, 13 and 14.

D.3 RTP-LX Safety Evaluation

We also do a safety evaluation study for the Hindi models using the RTP-LX (de Wynter et al., 2024) Hindi dataset. Only LLM evaluators were used for this study. We used the same instruction prompt as used in direct assessment above (Figure 11 and calculate the problematic content score in the model output generations. The problematic content rubric can be seen in Figure 15.

```
"""
# Role
You are an impartial judge and your task is to fairly evaluate the quality of the two responses provided for the question given below. The question and two responses are in {language}. You must choose the response that follows the provided guidelines and answers the question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, linguistic acceptability for {language}, and the level of detail of the responses. You must always provide a justification in English before your verdict. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor names of the responses. Be as objective as possible. You must follow the below provided verdict options and JSON format for your output.

## Verdict Options
"A" if response A is better than response B,
"B" if response B is better than response A,
"C" if both response A and response B are bad or equally good

## Output Format
{output_format}
"""

"""
## QUESTION
{prompt}

## Response A
{response_a}

## Response B
{response_b}
"""
```

Figure 10: LLM Pairwise Evaluation prompt

E Human Evaluation Setup

We employ an ethical data annotation company, KARYA to perform the pairwise evaluations as well as direct assessments. However, we do not engage them to do the safety evaluations due to ethical concerns. All annotators go through a training and screening check to maintain task performance. The task images displayed to the final annotators on smartphone screen are shown below.

E.1 Pairwise

For the pairwise evaluation, the annotators are shown a prompt as well as two responses in a fashion similar to the LLM. The annotation guidelines are given in Figure 16. The app interface for Hindi evaluation can be seen in Figure 17.

E.2 Direct Assessment

For the direct assessment, the annotators are shown the query-response pair. Then a flag is shown asking if the output is gibberish. If selected the response is given an automatic lowest score, otherwise, the annotators are asked to label the three metrics. The annotation guidelines are given in Figure 18. The app interface for Hindi evaluation can be seen in Figure 19.

Model	Short Name	Model Type
<i>Hindi Models</i>		
ai4bharat/Airavata (Gala et al., 2024)	Airavata	Indic
BhabhaAI/Gajendra-v0.1	Gajendra	Indic
GenVRadmin/Llamavaad	Llamavaad	Indic
manishiitg/open-aditi-hi-v4	Open-Aditi	Indic
GenVRadmin/AryaBhatta-GemmaGenZ-Vikas-Merged	AryaBhatta-GemmaGenZ	Indic
CohereForAI/aya-23-35B (Aryabumi et al., 2024)	Aya-23 35B	Open-Source
<i>Tamil Models</i>		
abhinand/tamil-llama-7b-instruct-v0.2 (Balachandran, 2023)	abhinand-Tamil	Indic
<i>Telugu Models</i>		
abhinand/telugu-llama-7b-instruct-v0.1 (Balachandran, 2023)	abhinand-Telugu	Indic
Telugu-LLM-Labs/Telugu-Llama2-7B-v0-Instruct	TLL-Telugu	Indic
<i>Malayalam Models</i>		
abhinand/malayalam-llama-7b-instruct-v0.1 (Balachandran, 2023)	abhinand-Malayalam	Indic
VishnuPJ/MalayaLLM_7B_Instruct_v0.2	MalayaLLM	Indic
<i>Kannada Models</i>		
Tensoic/Kan-Llama-7B-SFT-v0.5	Kan-Llama	Indic
Cognitive-Lab/Ambari-7B-Instruct-v0.1	Ambari	Indic
<i>Bengali Models</i>		
OdiaGenAI/odiagenAI-bengali-base-model-v1 (Parida et al., 2023)	OdiaGenAI-Bengali	Indic
<i>Odia Models</i>		
OdiaGenAI/odia_llama2_7B_base (Parida et al., 2023)	OdiaGenAI-Odia	Indic
<i>Marathi Models</i>		
smallstepai/Misal-7B-instruct-v0.1	Misal	Indic

Table 2: Details for models evaluated only on single languages.

F Evaluation Statistics

We perform a total of over 90k human evaluations and 30k LLM evaluations in our study. A breakdown of these statistics can be seen in Table 4. A total of 21690 datapoints (battles) are evaluated by LLMs evaluator for pairwise evaluation whereas 2880 model query-response pairs are evaluated for 3 metrics (hallucinations, task quality and linguistic acceptability) which results to 8640 datapoints in total. Hence, a total of $21690 + 8640 = 30330$ evaluations are performed by the LLM evaluator and since we annotate each datapoint by 3 humans, a total of 90990 evaluations are done by humans. In the models column, first number within parenthesis is the number of Indic-only models and the second value is the number of multilingual models under evaluation. We evaluate 12-15 models for each language (except Hindi) and 20 models for

Hindi.

G Leaderboards

In this section we present the detailed leaderboards constructed by the strategies discussed in Section 3. To calculate the Elo rating, we bootstrap the upsampled data 100 times. This is done due to the lower number of datapoints and to get confidence intervals.

G.1 MLE Elo Leaderboards

We report the MLE Elo leaderboards for all the 10 languages in Tables 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14.

G.2 Standard Elo Leaderboards

We report the Standard Elo leaderboards for all the 10 languages in Tables 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24.

Model	Short Name	Model Type
<i>OpenAI Models</i>		
gpt-4o (OpenAI et al., 2024)	GPT-4o	Proprietary
gpt-4 (OpenAI et al., 2024)	GPT-4	Proprietary
gpt-35-turbo (Brown et al., 2020)	GPT-35-Turbo	Proprietary
<i>Meta Models</i>		
meta-llama/Llama-2-7b-chat-hf (Touvron et al., 2023)	Llama-2 7B	Open-Source
meta-llama/Meta-Llama-3-8B-Instruct (AI@Meta, 2024)	Llama-3 8B	Open-Source
meta-llama/Meta-Llama-3-70B-Instruct (AI@Meta, 2024)	Llama-3 70B	Open-Source
<i>Google Models</i>		
gemini-pro [†] (Anil and Team, 2024)	Gemini-Pro 1.0	Proprietary
gemma-7b-it (Mesnard and Team, 2024)	Gemma 7B	Open-Source
<i>Mistral Models</i>		
mistralai/Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Mistral 7B	Open-Source
<i>Indic Models</i>		
GenVRadmin/AryaBhatta-GemmaOrca-Merged ^{††}	AryaBhatta-GemmaOrca	Open-Source
GenVRadmin/AryaBhatta-GemmaUltra-Merged ^{††}	AryaBhatta-GemmaUltra	Open-Source
GenVRadmin/llama38bGenZ_Vikas-Merged	AryaBhatta-Llama3GenZ	Open-Source
Telugu-LLM-Labs/Indic-gemma-7b-finetuned-sft-Navarasa-2.0	Navarasa	Open-Source
SamwaadLLM ^{†††}	SamwaadLLM	Open-Source

Table 3: Details for models evaluated on multiple languages. [†]Only Hindi and Bengali. ^{††}All languages except Marathi. ^{†††}All languages except Kannada and Malayalam.

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. Both LLM evaluator and Humans were used for all datapoints. In the models column, first number within parenthesis is the number of Indic-only models and the second value is the number of multilingual models under evaluation. Total evaluations: $21690 + 8640 = 30330$ for LLM, and $3 \times 30330 = 90990$ for humans, as each data point was annotated by 3 humans.

G.3 Direct Assessment Leaderboards

We report the Direct Assessment leaderboards for all the 10 languages in Tables 25, 26, 27, 28, 29, 30, 31, 32, 33 and 34.

H Detailed Agreement Scores

H.1 Percentage Agreement

In this section we report the Percentage Agreement (PA) scores which gives a raw-interpretable number but does not take class-imbalance into account. The scores are reported in Table 35.

Pairwise Battles On average humans agree on 70% of the samples among themselves whereas the accuracy is similar albeit slightly lower for humans and LLM evaluator. We see both evaluators agree more on the non-cultural subset of queries and follow a similar trend to the Fleiss Kappa correlations reported in Table 1. A language-wise breakdown of PA scores can be seen in Figure 20. We note that humans and LLM evaluator tend to agree less on Marathi, Punjabi and Bengali.

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1551 \pm 18.95	3	1604 \pm 22.73
Llama-3 70B	2	1444 \pm 13.09	5	1538 \pm 18.09
Gemini-Pro 1.0	3	1444 \pm 15.93	2	1672 \pm 21.87
GPT-4	4	1346 \pm 12.59	4	1598 \pm 20.44
SamwaadLLM	5	1247 \pm 11.98	1	1688 \pm 21.51
Llama-3 8B	6	1116 \pm 12.37	6	1233 \pm 16.0
Navarasa	7	1095 \pm 12.27	11	955 \pm 12.85
AryaBhatta-GemmaOrca	8	1067 \pm 10.7	10	975 \pm 12.91
AryaBhatta-Llama3GenZ	9	1066 \pm 10.17	7	1157 \pm 14.33
GPT-3.5-Turbo	10	1053 \pm 10.71	8	1086 \pm 13.49
AryaBhatta-GemmaUltra	11	1025 \pm 10.88	12	935 \pm 13.08
OdiaGenAI-Bengali	12	860 \pm 9.39	15	719 \pm 11.09
Gemma 7B	13	859 \pm 9.29	9	1029 \pm 14.42
Mistral 7B	14	821 \pm 8.97	13	891 \pm 12.85
Llama-2 7B	15	800 \pm 0.0	14	800 \pm 0.0

Table 5: MLE Elo for Bengali

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1399 \pm 15.59	1	1787 \pm 25.14
Llama-3 70B	2	1360 \pm 13.1	3	1704 \pm 22.18
GPT-4	3	1286 \pm 11.68	4	1675 \pm 20.88
SamwaadLLM	4	1246 \pm 11.78	2	1748 \pm 23.56
AryaBhatta-Llama3GenZ	5	1126 \pm 10.1	5	1441 \pm 19.08
Navarasa	6	1113 \pm 12.37	7	1237 \pm 19.66
AryaBhatta-GemmaOrca	7	1108 \pm 11.7	6	1285 \pm 21.41
AryaBhatta-GemmaUltra	8	1061 \pm 10.21	10	1175 \pm 19.59
GPT-3.5-Turbo	9	1042 \pm 11.21	9	1223 \pm 18.16
Llama-3 8B	10	995 \pm 9.55	8	1235 \pm 18.73
Gemma 7B	11	815 \pm 8.83	11	1028 \pm 16.83
Llama-2 7B	12	800 \pm 0.0	12	800 \pm 0.0
Mistral 7B	13	797 \pm 8.24	13	747 \pm 12.67

Table 6: MLE Elo for Gujarati

Direct Assessment For this task, we find slightly higher agreement between humans in comparison to the pairwise evaluation and it is similar for both the *query_types*. However we see a decline between human and LLM evaluator agreement and it is the worse for culturally-nuanced set of queries. From Figure 20, we again find the lowest agreement on Odia and Bengali for humans and LLM evaluator, similar to the Fleiss Kappa scores.

for the pairwise evaluation. Gujarati has the highest correlation while Bengali is the lowest. For the direct assessment the scores drops and it is again the lowest for Bengali. We conclude that evaluators are better at capturing general trends in a pairwise comparison evaluation over a direct assessment task.

H.2 Kendall Tau

We also report a detailed language-wise breakdown of the Kendall Tau rank correlations for both evaluation tasks in Table 36. We find the correlation between human and LLM leaderboards to be higher

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1607 \pm 16.12	1	1769 \pm 20.48
Aya-23 35B	2	1549 \pm 14.69	3	1597 \pm 16.51
SamwaadLLM	3	1521 \pm 14.49	4	1575 \pm 18.22
Llama-3 70B	4	1457 \pm 10.97	6	1440 \pm 14.49
Gemini-Pro 1.0	5	1454 \pm 12.79	2	1618 \pm 18.73
GPT-4	6	1407 \pm 13.03	5	1446 \pm 15.92
AryaBhatta-GemmaOrca	7	1278 \pm 12.07	11	1169 \pm 14.37
AryaBhatta-GemmaUltra	8	1260 \pm 12.4	10	1172 \pm 13.96
Navarasa	9	1259 \pm 12.59	9	1192 \pm 14.48
AryaBhatta-Llama3GenZ	10	1225 \pm 10.79	7	1240 \pm 13.45
AryaBhatta-GemmaGenZ	11	1205 \pm 11.82	14	1065 \pm 14.4
Llama-3 8B	12	1177 \pm 10.64	12	1161 \pm 13.64
Llamavaad	13	1169 \pm 12.2	8	1238 \pm 15.17
Gajendra	14	1158 \pm 9.78	13	1153 \pm 15.76
Airavata	15	1129 \pm 11.95	17	996 \pm 14.63
Gemma 7B	16	1070 \pm 11.79	15	1034 \pm 12.62
GPT-3.5-Turbo	17	1024 \pm 12.76	16	996 \pm 14.75
Open-Aditi	18	944 \pm 11.24	18	939 \pm 13.36
Mistral 7B	19	921 \pm 11.98	19	830 \pm 14.48
Llama-2 7B	20	800 \pm 0.0	20	800 \pm 0.0

Table 7: MLE Elo for Hindi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1420 \pm 18.35	2	1571 \pm 18.88
AryaBhatta-GemmaOrca	2	1406 \pm 18.03	5	1465 \pm 19.95
AryaBhatta-GemmaUltra	3	1395 \pm 15.7	4	1520 \pm 19.85
GPT-4o	4	1337 \pm 16.62	1	1676 \pm 18.78
GPT-4	5	1328 \pm 17.52	3	1560 \pm 17.8
Kan-Llama	6	1286 \pm 16.44	9	1298 \pm 17.18
Navarasa	7	1285 \pm 16.56	6	1379 \pm 16.73
AryaBhatta-Llama3GenZ	8	1261 \pm 15.03	7	1352 \pm 16.73
Ambari	9	1246 \pm 15.25	11	1218 \pm 16.42
Llama-3 8B	10	1246 \pm 15.34	8	1331 \pm 16.02
GPT-3.5-Turbo	11	1162 \pm 15.08	10	1223 \pm 14.43
Gemma 7B	12	967 \pm 14.35	12	1088 \pm 15.25
Mistral 7B	13	847 \pm 16.92	13	864 \pm 15.21
Llama-2 7B	14	800 \pm 0.0	14	800 \pm 0.0

Table 8: MLE Elo for Kannada

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1332 \pm 13.5	1	1777 \pm 21.68
Llama-3 70B	2	1271 \pm 11.21	3	1484 \pm 16.7
AryaBhatta-GemmaOrca	3	1216 \pm 12.55	4	1361 \pm 16.75
GPT-4	4	1200 \pm 11.42	2	1660 \pm 23.11
Navarasa	5	1195 \pm 11.04	5	1299 \pm 17.24
AryaBhatta-GemmaUltra	6	1150 \pm 11.38	8	1246 \pm 17.02
abhinand-Malayalam	7	1134 \pm 10.64	7	1249 \pm 17.1
MalayaLLM	8	1082 \pm 9.65	10	1208 \pm 15.95
AryaBhatta-Llama3GenZ	9	1080 \pm 9.11	6	1261 \pm 14.73
Llama-3 8B	10	991 \pm 10.94	9	1209 \pm 14.03
GPT-3.5-Turbo	11	859 \pm 8.73	11	1078 \pm 15.92
Gemma 7B	12	831 \pm 8.0	12	975 \pm 15.71
Mistral 7B	13	819 \pm 7.65	14	788 \pm 13.46
Llama-2 7B	14	800 \pm 0.0	13	800 \pm 0.0

Table 9: MLE Elo for Malayalam

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1416 \pm 16.63	1	1845 \pm 24.03
Llama-3 70B	2	1279 \pm 15.11	3	1592 \pm 22.7
GPT-4	3	1138 \pm 9.34	2	1628 \pm 22.97
SamwaadLLM	4	1018 \pm 9.63	4	1458 \pm 22.44
Navarasa	5	994 \pm 8.76	5	1303 \pm 16.79
Llama-3 8B	6	929 \pm 8.98	6	1199 \pm 18.68
Misal	7	893 \pm 8.2	9	988 \pm 15.5
GPT-3.5-Turbo	8	865 \pm 7.3	7	1199 \pm 16.66
AryaBhatta-Llama3GenZ	9	828 \pm 7.01	10	922 \pm 17.13
Mistral 7B	10	808 \pm 6.36	11	890 \pm 14.68
Llama-2 7B	11	800 \pm 0.0	12	800 \pm 0.0
Gemma 7B	12	798 \pm 6.69	8	1033 \pm 16.48

Table 10: MLE Elo for Marathi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1371 \pm 14.76	1	1676 \pm 18.56
Llama-3 70B	2	1303 \pm 12.12	3	1429 \pm 15.77
Navarasa	3	1232 \pm 11.47	4	1313 \pm 16.07
AryaBhatta-GemmaOrca	4	1221 \pm 11.32	5	1312 \pm 16.51
AryaBhatta-GemmaUltra	5	1191 \pm 10.69	9	1220 \pm 14.25
GPT-4	6	1171 \pm 11.67	2	1516 \pm 14.56
AryaBhatta-Llama3GenZ	7	1084 \pm 9.49	8	1228 \pm 14.01
Llama-3 8B	8	1064 \pm 8.78	7	1244 \pm 13.12
SamwaadLLM	9	983 \pm 9.61	6	1250 \pm 14.18
GPT-3.5-Turbo	10	926 \pm 9.71	10	1180 \pm 13.17
OdiaGenAI-Odia	11	887 \pm 8.38	11	942 \pm 12.35
Llama-2 7B	12	800 \pm 0.0	12	800 \pm 0.0
Mistral 7B	13	796 \pm 7.44	13	799 \pm 10.55
Gemma 7B	14	780 \pm 8.14	14	633 \pm 12.43

Table 11: MLE Elo for Odia

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1315 \pm 13.65	1	1782 \pm 25.54
Llama-3 70B	2	1308 \pm 14.35	2	1736 \pm 22.23
GPT-4	3	1258 \pm 11.55	3	1725 \pm 21.35
Navarasa	4	1001 \pm 7.48	6	1351 \pm 17.74
AryaBhatta-GemmaUltra	5	996 \pm 9.27	10	1272 \pm 17.76
AryaBhatta-GemmaOrca	6	958 \pm 7.82	7	1311 \pm 18.26
SamwaadLLM	7	951 \pm 9.02	4	1460 \pm 19.91
GPT-3.5-Turbo	8	913 \pm 7.13	8	1307 \pm 18.67
Llama-3 8B	9	902 \pm 7.1	9	1301 \pm 18.22
AryaBhatta-Llama3GenZ	10	892 \pm 8.46	5	1384 \pm 18.88
Gemma 7B	11	807 \pm 6.05	11	1018 \pm 14.37
Mistral 7B	12	804 \pm 6.81	13	777 \pm 14.38
Llama-2 7B	13	800 \pm 0.0	12	800 \pm 0.0

Table 12: MLE Elo for Punjabi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1342 \pm 11.52	5	1520 \pm 19.02
GPT-4o	2	1287 \pm 12.37	1	1703 \pm 21.88
AryaBhatta-GemmaOrca	3	1271 \pm 10.5	4	1531 \pm 21.17
AryaBhatta-GemmaUltra	4	1258 \pm 12.15	7	1478 \pm 21.58
Navarasa	5	1221 \pm 9.7	3	1541 \pm 22.22
GPT-4	6	1176 \pm 9.67	6	1519 \pm 19.36
AryaBhatta-Llama3GenZ	7	1142 \pm 11.18	8	1377 \pm 19.37
abhinand-Tamil	8	1126 \pm 10.09	2	1559 \pm 21.2
SamwaadLLM	9	1054 \pm 9.53	9	1362 \pm 20.22
Llama-3 8B	10	1043 \pm 10.39	10	1177 \pm 18.64
Gemma 7B	11	940 \pm 9.61	11	1166 \pm 18.55
GPT-3.5-Turbo	12	932 \pm 8.9	12	1126 \pm 17.95
Mistral 7B	13	819 \pm 9.41	14	697 \pm 13.77
Llama-2 7B	14	800 \pm 0.0	13	800 \pm 0.0

Table 13: MLE Elo for Tamil

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1313 \pm 11.74	3	1565 \pm 17.29
GPT-4o	2	1294 \pm 12.35	2	1625 \pm 17.26
AryaBhatta-GemmaOrca	3	1276 \pm 12.74	4	1515 \pm 15.96
AryaBhatta-GemmaUltra	4	1258 \pm 12.96	6	1492 \pm 16.83
Navarasa	5	1184 \pm 11.61	5	1503 \pm 16.93
GPT-4	6	1154 \pm 9.98	1	1634 \pm 17.24
Llama-3 8B	7	1100 \pm 11.59	10	1336 \pm 14.95
AryaBhatta-Llama3GenZ	8	1089 \pm 10.07	8	1383 \pm 12.92
SamwaadLLM	9	1074 \pm 10.21	7	1433 \pm 15.88
abhinand-Telugu	10	1040 \pm 10.55	9	1341 \pm 17.27
GPT-3.5-Turbo	11	834 \pm 8.12	12	1193 \pm 15.14
Llama-2 7B	12	800 \pm 0.0	14	800 \pm 0.0
TLL-Telugu	13	798 \pm 7.47	13	868 \pm 10.85
Mistral 7B	14	784 \pm 6.67	15	785 \pm 10.3
Gemma 7B	15	784 \pm 7.11	11	1261 \pm 16.11

Table 14: MLE Elo for Telugu

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1522 \pm 21.79	3	1499 \pm 22.77
Llama-3 70B	2	1422 \pm 17.19	5	1444 \pm 19.07
Gemini-Pro 1.0	3	1420 \pm 21.5	2	1557 \pm 18.81
GPT-4	4	1328 \pm 19.92	4	1493 \pm 23.95
SamwaadLLM	5	1233 \pm 17.95	1	1569 \pm 22.26
Llama-3 8B	6	1107 \pm 18.05	6	1194 \pm 19.26
Navarasa	7	1088 \pm 20.58	11	939 \pm 19.32
AryaBhatta-GemmaOrca	8	1061 \pm 18.7	10	957 \pm 18.27
AryaBhatta-Llama3GenZ	9	1057 \pm 18.08	7	1126 \pm 22.24
GPT-3.5-Turbo	10	1046 \pm 17.03	8	1065 \pm 21.33
AryaBhatta-GemmaUltra	11	1020 \pm 17.28	12	920 \pm 21.15
Gemma 7B	12	860 \pm 14.8	9	1006 \pm 20.86
OdiaGenAI-Bengali	13	859 \pm 16.36	15	723 \pm 14.98
Mistral 7B	14	820 \pm 13.87	13	881 \pm 18.95
Llama-2 7B	15	800 \pm 0.0	14	800 \pm 0.0

Table 15: Standard Elo for Bengali

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1376 \pm 18.87	1	1639 \pm 21.85
Llama-3 70B	2	1345 \pm 19.92	3	1564 \pm 23.01
GPT-4	3	1270 \pm 21.05	4	1546 \pm 19.72
SamwaadLLM	4	1233 \pm 19.59	2	1603 \pm 22.42
AryaBhatta-Llama3GenZ	5	1114 \pm 19.89	5	1346 \pm 20.42
Navarasa	6	1106 \pm 20.95	7	1167 \pm 21.8
AryaBhatta-GemmaOrca	7	1097 \pm 18.54	6	1209 \pm 21.52
AryaBhatta-GemmaUltra	8	1055 \pm 18.39	10	1111 \pm 23.22
GPT-3.5-Turbo	9	1034 \pm 18.07	9	1157 \pm 20.09
Llama-3 8B	10	986 \pm 16.73	8	1162 \pm 18.46
Gemma 7B	11	813 \pm 12.8	11	984 \pm 20.48
Llama-2 7B	12	800 \pm 0.0	12	800 \pm 0.0
Mistral 7B	13	796 \pm 14.02	13	760 \pm 14.24

Table 16: Standard Elo for Gujarati

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1603 \pm 23.02	1	1726 \pm 25.91
Aya-23 35B	2	1542 \pm 21.56	3	1573 \pm 27.42
SamwaadLLM	3	1516 \pm 22.23	4	1553 \pm 28.24
Llama-3 70B	4	1451 \pm 19.06	6	1424 \pm 27.02
Gemini-Pro 1.0	5	1450 \pm 17.85	2	1589 \pm 23.39
GPT-4	6	1402 \pm 24.63	5	1427 \pm 27.31
AryaBhatta-GemmaOrca	7	1276 \pm 24.56	11	1161 \pm 28.61
AryaBhatta-GemmaUltra	8	1256 \pm 20.98	10	1164 \pm 24.17
Navarasa	9	1254 \pm 21.93	9	1186 \pm 25.47
AryaBhatta-Llama3GenZ	10	1218 \pm 20.03	7	1230 \pm 27.51
AryaBhatta-GemmaGenZ	11	1203 \pm 25.32	14	1056 \pm 27.45
Llama-3 8B	12	1172 \pm 21.69	12	1151 \pm 24.99
Llamavaad	13	1167 \pm 22.82	8	1228 \pm 25.17
Gajendra	14	1152 \pm 23.01	13	1143 \pm 29.98
Airavata	15	1123 \pm 25.65	16	989 \pm 24.15
Gemma 7B	16	1069 \pm 19.8	15	1025 \pm 26.48
GPT-3.5-Turbo	17	1021 \pm 21.56	17	986 \pm 23.78
Open-Aditi	18	942 \pm 20.23	18	934 \pm 25.16
Mistral 7B	19	919 \pm 19.46	19	830 \pm 25.48
Llama-2 7B	20	800 \pm 0.0	20	800 \pm 0.0

Table 17: Standard Elo for Hindi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1397 \pm 20.42	2	1505 \pm 18.1
AryaBhatta-GemmaOrca	2	1380 \pm 20.99	5	1403 \pm 20.2
AryaBhatta-GemmaUltra	3	1374 \pm 19.9	4	1453 \pm 20.5
GPT-4o	4	1313 \pm 24.06	1	1608 \pm 17.34
GPT-4	5	1308 \pm 20.05	3	1498 \pm 22.13
Kan-Llama	6	1267 \pm 22.14	9	1241 \pm 22.49
Navarasa	7	1261 \pm 20.57	6	1319 \pm 18.66
AryaBhatta-Llama3GenZ	8	1237 \pm 20.92	7	1294 \pm 17.35
Llama-3 8B	9	1228 \pm 20.21	8	1273 \pm 19.28
Ambari	10	1224 \pm 24.8	11	1161 \pm 18.28
GPT-3.5-Turbo	11	1139 \pm 18.66	10	1169 \pm 16.92
Gemma 7B	12	952 \pm 22.18	12	1041 \pm 15.84
Mistral 7B	13	842 \pm 19.4	13	849 \pm 14.37
Llama-2 7B	14	800 \pm 0.0	14	800 \pm 0.0

Table 18: Standard Elo for Kannada

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1323 \pm 18.4	1	1680 \pm 17.7
Llama-3 70B	2	1268 \pm 18.7	3	1434 \pm 18.53
AryaBhatta-GemmaOrca	3	1210 \pm 18.82	4	1316 \pm 20.12
GPT-4	4	1196 \pm 22.36	2	1583 \pm 21.49
Navarasa	5	1190 \pm 19.39	5	1257 \pm 21.54
AryaBhatta-GemmaUltra	6	1144 \pm 18.17	8	1204 \pm 21.51
abhinand-Malayalam	7	1128 \pm 19.04	7	1210 \pm 21.92
MalayaLLM	8	1080 \pm 17.55	9	1172 \pm 20.09
AryaBhatta-Llama3GenZ	9	1077 \pm 17.8	6	1220 \pm 19.14
Llama-3 8B	10	988 \pm 15.11	10	1171 \pm 22.18
GPT-3.5-Turbo	11	858 \pm 15.46	11	1049 \pm 19.69
Gemma 7B	12	831 \pm 12.8	12	954 \pm 19.98
Mistral 7B	13	820 \pm 12.67	14	786 \pm 14.29
Llama-2 7B	14	800 \pm 0.0	13	800 \pm 0.0

Table 19: Standard Elo for Malayalam

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1376 \pm 16.72	1	1665 \pm 17.18
Llama-3 70B	2	1260 \pm 18.03	3	1480 \pm 20.19
GPT-4	3	1132 \pm 14.33	2	1504 \pm 22.98
SamwaadLLM	4	1013 \pm 17.09	4	1371 \pm 22.63
Navarasa	5	990 \pm 15.86	5	1244 \pm 21.12
Llama-3 8B	6	928 \pm 14.57	7	1153 \pm 20.71
Misal	7	893 \pm 13.54	9	966 \pm 20.62
GPT-3.5-Turbo	8	865 \pm 11.12	6	1153 \pm 22.55
AryaBhatta-Llama3GenZ	9	830 \pm 11.44	10	907 \pm 20.93
Mistral 7B	10	809 \pm 9.27	11	875 \pm 18.33
Llama-2 7B	11	800 \pm 0.0	12	800 \pm 0.0
Gemma 7B	12	798 \pm 11.46	8	1005 \pm 20.05

Table 20: Standard Elo for Marathi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1359 \pm 17.41	1	1591 \pm 20.27
Llama-3 70B	2	1297 \pm 19.04	3	1374 \pm 21.94
Navarasa	3	1225 \pm 19.19	5	1260 \pm 24.21
AryaBhatta-GemmaOrca	4	1216 \pm 17.32	4	1264 \pm 18.88
AryaBhatta-GemmaUltra	5	1182 \pm 17.59	9	1176 \pm 18.41
GPT-4	6	1167 \pm 20.56	2	1454 \pm 21.35
AryaBhatta-Llama3GenZ	7	1083 \pm 17.02	8	1184 \pm 19.87
Llama-3 8B	8	1059 \pm 16.45	7	1196 \pm 18.82
SamwaadLLM	9	978 \pm 18.78	6	1206 \pm 20.07
GPT-3.5-Turbo	10	925 \pm 15.03	10	1136 \pm 16.91
OdiaGenAI-Odia	11	886 \pm 14.63	11	919 \pm 16.64
Llama-2 7B	12	800 \pm 0.0	12	800 \pm 0.0
Mistral 7B	13	798 \pm 12.32	13	797 \pm 13.61
Gemma 7B	14	781 \pm 12.07	14	664 \pm 16.39

Table 21: Standard Elo for Odia

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
GPT-4o	1	1299 \pm 16.33	1	1649 \pm 19.5
Llama-3 70B	2	1289 \pm 18.68	2	1606 \pm 16.87
GPT-4	3	1244 \pm 15.88	3	1597 \pm 20.04
Navarasa	4	998 \pm 15.47	6	1259 \pm 20.23
AryaBhatta-GemmaUltra	5	994 \pm 15.39	10	1187 \pm 18.53
AryaBhatta-GemmaOrca	6	958 \pm 15.62	7	1226 \pm 16.99
SamwaadLLM	7	947 \pm 11.94	4	1360 \pm 18.73
GPT-3.5-Turbo	8	910 \pm 13.84	8	1219 \pm 18.26
Llama-3 8B	9	901 \pm 14.53	9	1215 \pm 17.32
AryaBhatta-Llama3GenZ	10	890 \pm 14.87	5	1293 \pm 20.79
Gemma 7B	11	806 \pm 9.78	11	969 \pm 16.46
Mistral 7B	12	803 \pm 10.7	13	782 \pm 13.56
Llama-2 7B	13	800 \pm 0.0	12	800 \pm 0.0

Table 22: Standard Elo for Punjabi

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1333 \pm 18.39	5	1424 \pm 20.59
GPT-4o	2	1279 \pm 17.89	1	1598 \pm 21.51
AryaBhatta-GemmaOrca	3	1264 \pm 17.87	4	1433 \pm 21.65
AryaBhatta-GemmaUltra	4	1252 \pm 17.35	7	1381 \pm 21.87
Navarasa	5	1213 \pm 15.35	3	1445 \pm 19.92
GPT-4	6	1168 \pm 15.29	6	1419 \pm 23.23
AryaBhatta-Llama3GenZ	7	1136 \pm 17.26	8	1287 \pm 21.44
abhinand-Tamil	8	1118 \pm 16.55	2	1461 \pm 19.92
SamwaadLLM	9	1050 \pm 18.39	9	1274 \pm 20.81
Llama-3 8B	10	1037 \pm 16.76	10	1096 \pm 20.83
Gemma 7B	11	935 \pm 16.55	11	1090 \pm 17.9
GPT-3.5-Turbo	12	926 \pm 16.19	12	1059 \pm 18.81
Mistral 7B	13	817 \pm 13.25	14	730 \pm 12.24
Llama-2 7B	14	800 \pm 0.0	13	800 \pm 0.0

Table 23: Standard Elo for Tamil

Model	Rank (Human)	Elo Rating (Human)	Rank (LLM)	Elo Rating (LLM)
Llama-3 70B	1	1307 \pm 19.88	3	1483 \pm 19.91
GPT-4o	2	1283 \pm 21.23	2	1542 \pm 18.22
AryaBhatta-GemmaOrca	3	1267 \pm 16.56	4	1430 \pm 21.74
AryaBhatta-GemmaUltra	4	1252 \pm 20.73	6	1408 \pm 20.34
Navarasa	5	1177 \pm 19.01	5	1421 \pm 20.01
GPT-4	6	1145 \pm 17.17	1	1542 \pm 18.08
Llama-3 8B	7	1095 \pm 16.57	10	1258 \pm 17.49
AryaBhatta-Llama3GenZ	8	1080 \pm 17.91	8	1305 \pm 17.75
SamwaadLLM	9	1067 \pm 18.45	7	1355 \pm 20.9
abhinand-Telugu	10	1037 \pm 17.82	9	1266 \pm 21.08
GPT-3.5-Turbo	11	831 \pm 13.47	12	1122 \pm 16.63
Llama-2 7B	12	800 \pm 0.0	14	800 \pm 0.0
TLL-Telugu	13	796 \pm 10.8	13	857 \pm 14.31
Mistral 7B	14	784 \pm 9.63	15	788 \pm 11.95
Gemma 7B	15	784 \pm 11.21	11	1189 \pm 19.98

Table 24: Standard Elo for Telugu

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
GPT-4o	1	1.80	1.75	0.80	4.35	1	2	2	1	5
Llama-3 70B	2	1.70	1.60	0.75	4.05	1	2	2	1	5
Gemini-Pro 1.0	3	1.35	1.40	0.60	3.35	1	2	2	1	5
AryaBhatta-GemmaOrca	4	1.15	0.85	0.60	2.60	10	1.20	1.35	0.65	3.20
Navarasa	5	1.05	0.90	0.55	2.50	12	1	1.15	0.65	2.80
AryaBhatta-GemmaUltra	6	1.15	0.80	0.50	2.45	11	1.10	1.15	0.65	2.90
GPT-3.5-Turbo	7	1.05	0.90	0.30	2.25	7	1.90	1.95	0.95	4.80
Llama-3 8B	8	1.25	0.70	0.30	2.25	8	1.85	1.85	0.95	4.65
GPT-4	9	0.80	0.95	0.40	2.15	1	2	2	1	5
SamwaadLLM	10	0.85	0.80	0.40	2.05	1	2	2	1	5
AryaBhatta-Llama3GenZ	11	0.80	0.70	0.40	1.90	6	2	1.95	1	4.95
Gemma 7B	12	0.25	0.10	0.05	0.40	9	1.70	1.75	0.80	4.25
OdiaGenAI-Bengali	13	0.25	0.05	0.05	0.35	14	0.25	0.20	0.05	0.50
Mistral 7B	14	0.05	0.05	0	0.10	13	1.25	1.15	0.35	2.75
Llama-2 7B	15	0	0	0	0	15	0.10	0.05	0	0.15

Table 25: Direct Assessment Leaderboard for Bengali

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.90	1.60	0.75	4.25	1	2	2	1	5
GPT-4o	2	1.40	1.70	0.65	3.75	1	2	2	1	5
AryaBhatta-Llama3GenZ	3	1.70	1	0.45	3.15	5	1.90	1.80	0.95	4.65
GPT-4	4	0.95	1.10	0.45	2.50	1	2	2	1	5
SamwaadLLM	5	1.10	0.95	0.35	2.40	1	2	2	1	5
AryaBhatta-GemmaOrca	6	1.15	0.70	0.40	2.25	8	1.30	1.20	0.70	3.20
AryaBhatta-GemmaUltra	7	1.05	0.60	0.30	1.95	9	1.20	1.30	0.55	3.05
Navarasa	8	0.90	0.60	0.30	1.80	10	1.30	1.20	0.55	3.05
GPT-3.5-Turbo	9	0.75	0.70	0.30	1.75	6	2	1.75	0.90	4.65
Llama-3 8B	10	1.10	0.45	0.15	1.70	7	1.95	1.70	0.80	4.45
Gemma 7B	11	0	0	0	0	11	0.45	0.95	0.40	1.80
Mistral 7B	12	0	0	0	0	12	0.10	0	0.05	0.15
Llama-2 7B	13	0	0	0	0	13	0	0	0.05	0.05

Table 26: Direct Assessment Leaderboard for Gujarati

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
GPT-4o	1	1.95	2	1	4.95	1	2	2	1	5
Llama-3 70B	2	1.95	1.95	0.95	4.85	1	2	2	1	5
Gemini-Pro 1.0	3	1.95	1.95	0.90	4.80	1	2	2	1	5
AryaBhatta-Llama3GenZ	4	1.90	1.90	0.90	4.70	1	2	2	1	5
GPT-3.5-Turbo	5	1.95	1.80	0.75	4.50	9	2	1.80	1	4.80
AryaBhatta-GemmaGenZ	6	2	1.60	0.60	4.20	15	1.55	1.55	0.95	4.05
AryaBhatta-GemmaOrca	7	2	1.60	0.60	4.20	11	1.70	1.80	0.95	4.45
SamwaadLLM	8	1.75	1.70	0.70	4.15	6	1.95	2	1	4.95
Aya-23 35B	9	1.90	1.65	0.60	4.15	1	2	2	1	5
GPT-4	10	1.75	1.65	0.70	4.10	7	2	1.95	1	4.95
Llama-3 8B	11	1.85	1.55	0.55	3.95	8	1.95	1.95	0.95	4.85
AryaBhatta-GemmaUltra	12	1.95	1.45	0.50	3.90	14	1.60	1.65	0.90	4.15
Navarasa	13	2	1.40	0.40	3.80	12	1.70	1.70	1	4.40
Gajendra	14	1.95	1.15	0.35	3.45	13	1.80	1.75	0.85	4.40
Airavata	15	1.85	0.90	0.15	2.90	19	1.20	1.20	0.50	2.90
Llamavaad	16	1.25	1	0	2.25	10	2	1.75	1	4.75
Gemma 7B	17	1	0.85	0.05	1.90	16	1.60	1.65	0.75	4
Open-Aditi	18	0.90	0.55	0	1.45	17	1.60	1.50	0.70	3.80
Mistral 7B	19	0.70	0.30	0	1	18	1.10	1.30	0.55	2.95
Llama-2 7B	20	0.50	0.10	0	0.60	20	0.45	0.40	0.20	1.05

Table 27: Direct Assessment Leaderboard for Hindi

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.95	1.50	0.80	4.25	1	2	2	1	5
Llama-3 8B	2	1.85	1.05	0.65	3.55	6	1.95	1.80	0.95	4.70
AryaBhatta-GemmaOrca	3	1.55	1.15	0.70	3.40	9	1.60	1.70	0.75	4.05
AryaBhatta-GemmaUltra	4	1.55	1.10	0.65	3.30	7	1.75	1.75	0.90	4.40
GPT-4o	5	1.35	1.30	0.50	3.15	1	2	2	1	5
AryaBhatta-Llama3GenZ	6	1.60	0.80	0.60	3	3	2	1.95	1	4.95
Navarasa	7	1.60	0.90	0.50	3	8	1.65	1.70	0.80	4.15
GPT-4	8	1.60	0.95	0.40	2.95	5	2	1.85	1	4.85
Ambari	9	1.55	0.85	0.45	2.85	10	1.45	1.25	0.55	3.25
Kan-Llama	10	1.50	0.65	0.30	2.45	11	1.35	1.20	0.70	3.25
GPT-3.5-Turbo	11	1.65	0.50	0.25	2.40	4	2	1.90	0.95	4.85
Gemma 7B	12	0.35	0.05	0.05	0.45	12	0.95	0.80	0.35	2.10
Llama-2 7B	13	0.45	0	0	0.45	14	0	0	0	0
Mistral 7B	14	0.30	0	0	0.30	13	0.45	0.10	0	0.55

Table 28: Direct Assessment Leaderboard for Kannada

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.95	1.50	0.70	4.15	1	2	2	1	5
Navarasa	2	1.65	1.15	0.60	3.40	4	1.85	1.80	1	4.65
AryaBhatta-GemmaOrca	3	1.65	1.15	0.60	3.40	6	1.80	1.80	0.90	4.50
GPT-4o	4	1.40	1.35	0.45	3.20	3	2	1.95	1	4.95
AryaBhatta-GemmaUltra	5	1.45	0.90	0.40	2.75	9	1.45	1.45	0.70	3.60
AryaBhatta-Llama3GenZ	6	1.30	0.65	0.45	2.40	5	1.85	1.80	0.95	4.60
Llama-3 8B	7	1.25	0.50	0.45	2.20	7	1.80	1.70	0.90	4.40
GPT-4	8	0.95	0.75	0.25	1.95	1	2	2	1	5
MalayaLLM	9	0.90	0.65	0.30	1.85	11	1.10	1.05	0.55	2.70
abhinand-Malayalam	10	0.95	0.60	0.25	1.80	10	1.10	1.25	0.55	2.90
GPT-3.5-Turbo	11	0.60	0.05	0.10	0.75	8	1.80	1.45	0.90	4.15
Gemma 7B	12	0.10	0.05	0.05	0.20	12	0.45	0.70	0.30	1.45
Mistral 7B	13	0	0	0	0	13	0.20	0.15	0.05	0.40
Llama-2 7B	14	0	0	0	0	14	0.10	0	0.15	0.25

Table 29: Direct Assessment Leaderboard for Malayalam

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
GPT-4o	1	1.90	1.90	0.90	4.70	1	2	2	1	5
Llama-3 70B	2	1.75	1.70	0.85	4.30	1	2	2	1	5
GPT-4	3	1.30	1.20	0.55	3.05	1	2	2	1	5
SamwaadLLM	4	1.70	0.85	0.45	3	5	2	1.75	0.75	4.50
Navarasa	5	1.55	0.85	0.45	2.85	6	1.70	1.75	0.85	4.30
GPT-3.5-Turbo	6	1.35	0.75	0.30	2.40	4	2	1.80	0.90	4.70
Misal	7	1.80	0.40	0.15	2.35	9	1.20	0.70	0.65	2.55
Llama-3 8B	8	1.15	0.65	0.30	2.10	7	1.65	1.60	0.80	4.05
Gemma 7B	9	0.20	0.15	0	0.35	8	1.20	1.35	0.60	3.15
AryaBhatta-Llama3GenZ	10	0.20	0	0	0.20	10	0.70	0.45	0.35	1.50
Llama-2 7B	11	0.05	0	0	0.05	12	0.30	0.10	0.10	0.50
Mistral 7B	12	0	0	0	0	11	0.85	0.35	0.10	1.30

Table 30: Direct Assessment Leaderboard for Marathi

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.35	1.30	0.65	3.30	1	2	2	1	5
GPT-4o	2	0.75	1.25	0.55	2.55	1	2	2	1	5
Navarasa	3	1.05	0.90	0.45	2.40	9	1.25	1.35	0.60	3.20
AryaBhatta-GemmaOrca	4	1	0.75	0.50	2.25	8	1.50	1.55	0.75	3.80
AryaBhatta-Llama3GenZ	5	0.70	0.55	0.35	1.60	7	1.80	1.50	0.70	4
GPT-4	6	0.30	0.85	0.35	1.50	4	2	1.90	1	4.90
AryaBhatta-GemmaUltra	7	0.70	0.55	0.20	1.45	10	1.15	0.95	0.55	2.65
Llama-3 8B	8	0.50	0.35	0.20	1.05	3	2	1.95	1	4.95
SamwaadLLM	9	0.25	0.30	0.10	0.65	6	1.55	1.60	0.90	4.05
OdiaGenAI-Odia	10	0.10	0.05	0.05	0.20	11	0.95	0.50	0.30	1.75
GPT-3.5-Turbo	11	0	0	0	0	5	1.90	1.80	0.90	4.60
Llama-2 7B	12	0	0	0	0	12	0.15	0	0.20	0.35
Mistral 7B	13	0	0	0	0	13	0.10	0	0	0.10
Gemma 7B	14	0	0	0	0	14	0	0	0	0

Table 31: Direct Assessment Leaderboard for Odia

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
GPT-4o	1	1.95	1.85	0.90	4.70	1	2	2	1	5
Llama-2 70B	2	2	1.70	0.75	4.45	1	2	2	1	5
GPT-4	3	1.75	1.55	0.75	4.05	1	2	2	1	5
AryaBhatta-GemmaUltra	4	1.95	1.05	0.40	3.40	9	1.60	1.35	0.70	3.65
Navarasa	5	1.85	0.85	0.40	3.10	8	1.65	1.50	0.70	3.85
AryaBhatta-GemmaOrca	6	1.65	0.85	0.30	2.80	10	1.65	1.35	0.60	3.60
AryaBhatta-Llama3GenZ	7	1.95	0.45	0.20	2.60	7	1.75	1.40	0.80	3.95
GPT-3.5-Turbo	8	1.55	0.70	0.30	2.55	4	2	1.65	0.90	4.55
Llama-3 8B	9	1.55	0.55	0.20	2.30	6	1.85	1.45	0.70	4
SamwaadLLM	10	1.10	0.55	0.30	1.95	5	1.85	1.60	0.75	4.20
Gemma 7B	11	0	0	0	0	11	0.40	0.70	0.10	1.20
Mistral 7B	12	0	0	0	0	12	0.10	0	0	0.10
Llama-2 7B	13	0	0	0	0	13	0	0	0	0

Table 32: Direct Assessment Leaderboard for Punjabi

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.90	1.75	1	4.65	3	2	1.95	1	4.95
Navarasa	2	1.85	1.45	0.80	4.10	5	2	1.75	0.95	4.70
AryaBhatta-GemmaOrca	3	1.75	1.15	0.70	3.60	7	1.80	1.85	0.90	4.55
GPT-4o	4	1.35	1.10	0.65	3.10	1	2	2	1	5
AryaBhatta-Llama3GenZ	5	1.40	0.95	0.70	3.05	4	2	1.80	1	4.80
AryaBhatta-GemmaUltra	6	1.50	1	0.50	3	10	1.55	1.60	0.85	4
abhinand-Tamil	7	1.55	0.80	0.55	2.90	6	1.95	1.80	0.90	4.65
Llama-3 8B	8	1.70	0.45	0.60	2.75	9	1.85	1.60	0.80	4.25
SamwaadLLM	9	1.25	0.55	0.55	2.35	8	1.90	1.60	0.75	4.25
GPT-4	10	0.90	0.65	0.45	2	1	2	2	1	5
GPT-3.5-Turbo	11	1	0.25	0.15	1.40	11	1.80	1.30	0.70	3.80
Gemma 7B	12	0.45	0.10	0.20	0.75	12	1.65	1.25	0.60	3.50
Llama-2 7B	13	0.10	0	0	0.10	13	0.40	0.15	0.05	0.60
Mistral 7B	14	0	0	0	0	14	0.15	0.05	0	0.20

Table 33: Direct Assessment Leaderboard for Tamil

Model	Rank (Human)	LA (Human)	TQ (Human)	H (Human)	Score (Human)	Rank (LLM)	LA (LLM)	TQ (LLM)	H (LLM)	Score (LLM)
Llama-3 70B	1	1.95	1.90	1	4.85	1	2	2	1	5
GPT-4o	2	1.90	1.65	0.95	4.50	1	2	2	1	5
GPT-4	3	1.95	1.60	0.90	4.45	4	2	1.95	0.95	4.90
Llama-3 8B	4	1.90	1.40	0.90	4.20	1	2	2	1	5
Navarasa	5	1.80	1.45	0.90	4.15	7	1.85	1.80	0.90	4.55
AryaBhatta-Llama3GenZ	6	2	1.30	0.80	4.10	5	2	1.90	0.95	4.85
SamwaadLLM	7	1.90	1.30	0.80	4	6	2	1.90	0.95	4.85
AryaBhatta-GemmaOrca	8	1.70	1.45	0.80	3.95	8	1.75	1.75	0.85	4.35
AryaBhatta-GemmaUltra	9	1.70	1.45	0.80	3.95	9	1.75	1.75	0.85	4.35
GPT-3.5-Turbo	10	1.75	0.50	0.40	2.65	10	1.90	1.40	0.75	4.05
abhinand-Telugu	11	1.05	0.70	0.35	2.10	11	1.15	1.20	0.50	2.85
TLL-Telugu	12	1.05	0.05	0.05	1.15	13	0.50	0.25	0.10	0.85
Gemma 7B	13	0	0	0	0	12	1	1.05	0.45	2.50
Llama-2 7B	14	0	0	0	0	14	0.05	0.10	0.05	0.20
Mistral 7B	15	0	0	0	0	15	0.10	0.05	0	0.15

Table 34: Direct Assessment Leaderboard for Telugu

```

"""
# Role
You are a helpful assistant.

## Task
Question-Answering: Given a question and a response to that
question, your task is to evaluate the response with respect to
the given question and listed metric. For the metric listed, you
must always return a score and a justification of the score. Note
that, both the question and its response are given in language.
**Do not** allow the length of the response to influence your
evaluation.

### Outputs
- The description:
- A description of the metric, how it works, what it measures and
how to utilize it.

- The score:
- Scores are integer values in accordance to the metric description
provided.

- The justification:
- Justifications provide the evidence and step by step reasoning
on how the score is reached. Justifications must always be given
in English. Be as objective as possible.

- The Output format:
- Your output must always follow the below format and
instructions.
- {output_format}
"""

QUESTION = {question}
RESPONSE = {response}
LANGUAGE = {language}

Now, evaluate the above response in the context of the above given
question with regard to the following metric.

### Metric
You are given below the metric, with its description and scoring
schema in a JSON format.

```json
metric_description
```
"""

```

Figure 11: LLM Direct Assessment prompt

| Language | Pairwise | Direct |
|----------------|----------|--------|
| <i>Average</i> | 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 36: Kendall Tau (τ) correlations between Pairwise (Elo) and Direct Assessment leaderboards constructed through human annotators and LLM evaluator.

| Prompt Type | Pairwise | | Direct | |
|--------------|-------------------------------|--------------------|-------------------------------|--------------------|
| | \mathcal{H} - \mathcal{H} | \mathcal{H} -LLM | \mathcal{H} - \mathcal{H} | \mathcal{H} -LLM |
| <i>All</i> | 0.70 | 0.69 | 0.70 | 0.61 |
| Cultural | 0.67 | 0.65 | 0.71 | 0.57 |
| Non-Cultural | 0.73 | 0.73 | 0.70 | 0.65 |

Table 35: Average Percentage Agreement (PA) correlations between Humans and Human-LLM for both evaluations across prompt types. Here \mathcal{H} stands for Humans.

```

"name": "hallucinations",

"description": "Hallucinations assess the extent to which a model's output remains anchored to, and consistent with, the input content provided. Text with hallucinations while linguistically fluent, are factually baseless or counterfactual in relation to the input. These hallucinations can manifest as additions, omissions, or distortions, and might lead to outputs that are misleading or factually incorrect. This metric serves as a check against unwarranted deviations from the ground truth provided in the input. The scoring rubric is described below, with a few possible reasons (which might not be exhaustive) for a given score.",

"scoring": {
  "1": {
    "(a)": "The model's output is strictly aligned with and grounded in the information provided in the input.",
    "(b)": "No evidence of added, omitted, or distorted facts that weren't part of the original content.",
    "(c)": "Maintains the integrity of the original information without any unwarranted extrapolations."
  },
  "0": {
    "(a)": "The output introduces statements, claims, or details that weren't present or implied in the input.",
    "(b)": "Contains counterfactual information that directly conflicts with the input content.",
    "(c)": "Demonstrates unexplained deviations, extrapolations, or interpretations not grounded in the provided data."
  }
}
}

```

Figure 12: Metric description for complex instructions (Hallucinations).

```

"name": "task quality",

"description": "Task Quality gauges the degree to which a model adheres to and executes the specific directives given in the prompt. This metric zeroes in exclusively on the fidelity of the model's response to the prompt's instructions. An ideal response not only recognizes the overt commands of the prompt but also respects its nuance and subtleties. The scoring rubric is described below, with a few possible reasons (which might not be exhaustive) for a given score."

"scoring": {
  "0": {
    "(a)": "The model disregards the instructions entirely.",
    "(b)": "The output is entirely irrelevant to the prompt.",
    "(c)": "There is a clear disconnect between the user's request and the model's response."
  },
  "1": {
    "(a)": "The model grasps and addresses the main theme or element of the instruction but may miss out on finer details or nuances.",
    "(b)": "There is partial alignment with the prompt, indicating some elements of relevance, but not a complete match.",
    "(c)": "The response might include extraneous details not asked for, or it might omit some requested specifics."
  },
  "2": {
    "(a)": "The model demonstrates a precise understanding and adherence to the prompt's instructions.",
    "(b)": "The output holistically satisfies all aspects of the given directive without any deviation.",
    "(c)": "There's a clear and direct correlation between the user's instruction and the model's response, with no aspect of the instruction left unaddressed."
  }
}
}

```

Figure 13: Metric description for complex instructions (task quality).

```

"name": "linguistic_acceptability",

"description": "Linguistic acceptability pertains to the degree to which a given language structure (e.g., phrase, sentence, discourse) aligns with the implicit norms and rules of a native speaker's linguistic intuition. In the study of language, it's distinct from 'grammaticality', which is a stricter and narrower concept based on the prescriptive rules of a language. Linguistic acceptability, on the other hand, captures broader native-speaker intuitions and encompasses factors like fluency, idiomacy, and appropriateness in context. In the context of language models, evaluating linguistic acceptability involves assessing the output of the model not just for its adherence to grammar rules, but for its overall fit within the natural, expected, and intuitive contours of fluent human language. The scoring rubric is described below, with a few possible reasons (which might not be exhaustive) for a given score.",

"scoring": {
  "0": {
    "(a)": "Sentences that lack clear syntactic structure.",
    "(b)": "Usage of non-existent or incorrect words.",
    "(c)": "Grossly inappropriate word choices for a given context."
  },
  "1": {
    "(a)": "Overly verbose or stilted phrasing.",
    "(b)": "Minor grammatical errors that do not impede understanding.",
    "(c)": "Use of a word that's technically correct but not the most appropriate for context."
  },
  "2": {
    "(a)": "Seamless integration of contextually relevant vocabulary",
    "(b)": "Effective use of idiomatic expressions without sounding forced.",
    "(c)": "Sentences that reflect natural rhythm, emphasis, and intonation of spoken language."
  }
}
}

```

Figure 14: Metric description for complex instructions (Linguistic Acceptability).

```

"name": "problematic_content",

"description": "Problematic Content assesses the presence of any content that may be deemed offensive, inappropriate, or harmful in the model's output. This metric serves as a filter against outputs that might perpetuate harmful stereotypes, misinformation, or any form of bigotry. This metric highlights the importance of ethical considerations in outputs and underscores the need for vigilance against unintended harmful content. The scoring rubric is described below, with a few possible reasons (which might not be exhaustive) for a given score.",

"scoring": {
  "1": {
    "(a)": "The output is free from any content that might be perceived as offensive, harmful, or inappropriate.",
    "(b)": "No propagation of stereotypes, misinformation, or inappropriate assertions.",
    "(c)": "The response respects cultural, social, and individual sensitivities."
  },
  "0": {
    "(a)": "The output contains statements, suggestions, or themes that can be deemed problematic or offensive.",
    "(b)": "There is evidence of perpetuation of harmful stereotypes or biases.",
    "(c)": "Contains misinformation or promotes inappropriate or harmful narratives."
  }
}

```

Figure 15: Metric description for complex instructions (Problematic content).

Pairwise Evaluation

- You have to evaluate the responses/answers to the given prompt or question. Choose which is the best answer. Three options are available: response 1 is the best, response 2 is the best, and both are equal. Based on the option selected, you have to give a valid voice feedback of 20 seconds to 35 seconds.
- For example, if you had chosen response/answer 1 as the best, The feedback should be like, "A valid reason why you selected it and also explain why you have denied the other response."
- Don't add any extra information or facts about the response or answer, or try to explain the response or answer.
- Use the last option only if necessary: "Both are equal." If you are not able to analyse or both are blunders or garbage and not at all related to the prompt.
- If both answers/responses are good, select the very good answer/response out of the two.
- Out of the two responses, if the first one has 30-40% relevant data and content is repeated or some minor spelling mistakes and the second response is a blunder or garbage, we can select the first one. After selecting the first one, you can also include in the feedback audio that there are some minor spelling mistakes or incomplete data or repetition of the content. **This is applicable only if the second response is garbage or blunder or not at all related to prompt in anyway.**

Figure 16: Detailed task instructions provided to the annotators.

प्रत्येक कार्य में, आपको एक संकेत और दो अलग-अलग उत्तर दिखाए जाएंगे। उत्तरों को ध्यान से देखें और निर्णय लें कि कौन सा अच्छा है।

कवि कबीर दास ने भक्ति आंदोलन को महत्वपूर्ण रूप से प्रभावित किया। हिंदी भाषियों की सांस्कृतिक पहचान को आकार देने में उनकी हिंदी कविता के महत्व को स्पष्ट करें।

प्रतिक्रिया 1

कवि कबीर दास ने भक्ति आंदोलन को बहुत प्रभावित किया। हिंदी भाषियों की सांस्कृतिक पहचान को आकार देने में उनकी हिंदी कविता का महत्व बहुत है। उन्होंने हिंदी भाषा को एक नई पहचान दी, और इसे एक विशिष्ट भाषा के रूप में स्थापित किया। उनकी कविताएं हिंदी भाषा के ...

...See More

प्रतिक्रिया 2

कबीर दास की कविता ने भक्ति आंदोलन को महत्वपूर्ण रूप से प्रभावित किया और हिंदी भाषी लोगों की सांस्कृतिक पहचान को आकार देने में

प्रत्येक कार्य में, आपको एक संकेत और दो अलग-अलग उत्तर दिखाए जाएंगे। उत्तरों को ध्यान से देखें और निर्णय लें कि कौन सा अच्छा है।

और मानवता के महत्व को बल दिया। उनकी कविता सादी, स्पष्ट और आसानी से समझने वाली थी, जिससे यह सामान्य लोगों के लिए सुलभ हो गई। उनकी शिक्षाएं समाज के हर वर्ग के लोगों के साथ गूंजी, जिससे सामाजिक समानता और एकता को बढ़ावा मिला। उनकी कविता ने हिंदी साहित्य के विकास में भी महत्वपूर्ण भूमिका निभाई। उन्होंने हिंदी भाषा को साहित्यिक स्थिति में उन्नत किया और भाषा को और अधिक सम

कौन सा उत्तर सर्वोत्तम है?

☐ प्रतिक्रिया 1 बेहतर है

☐ प्रतिक्रिया 2 बेहतर है

☐ दोनों बराबर हैं

Figure 17: Hindi App screenshots for pairwise human evaluations

Direct Assessment

- Linguistic Acceptability
 - 0 : gibberish or bad sentence structure to native speaker.
 - 1 : some grammatical mistakes.
 - 2 : perfect language even if answer is incorrect. This should be irrespective of the answer it gave meaning if its hallucinated.
- Hallucinations
 - 0 : no stuff that is not factual.
 - 1 : even if there is one stuff that is not correct, gibberish also gets this.
- Task Quality
 - 0 : if its gibberish or totally hallucinated.
 - 1 : if it is partially correct like not giving all the details asked and only some, it is fine if there is little hallucination.
 - 2 : perfect answer, most of the things answered and no hallucinations.

Figure 18: Detailed task instructions provided to the annotators.

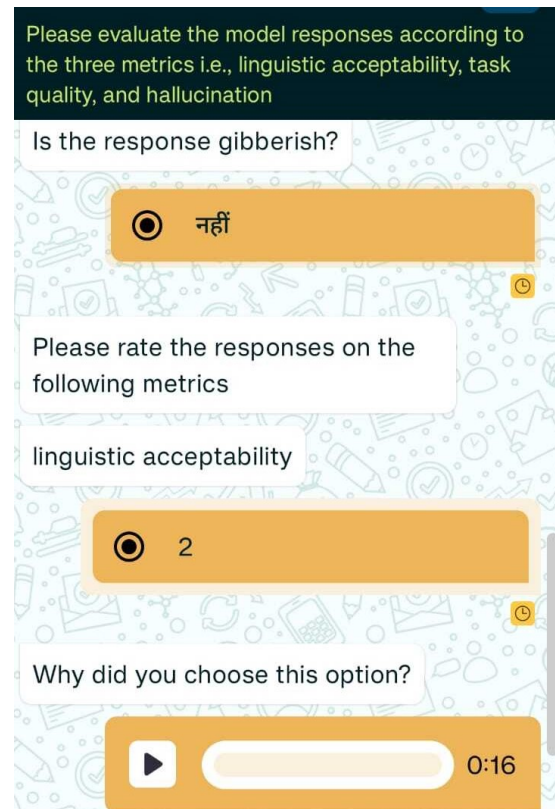
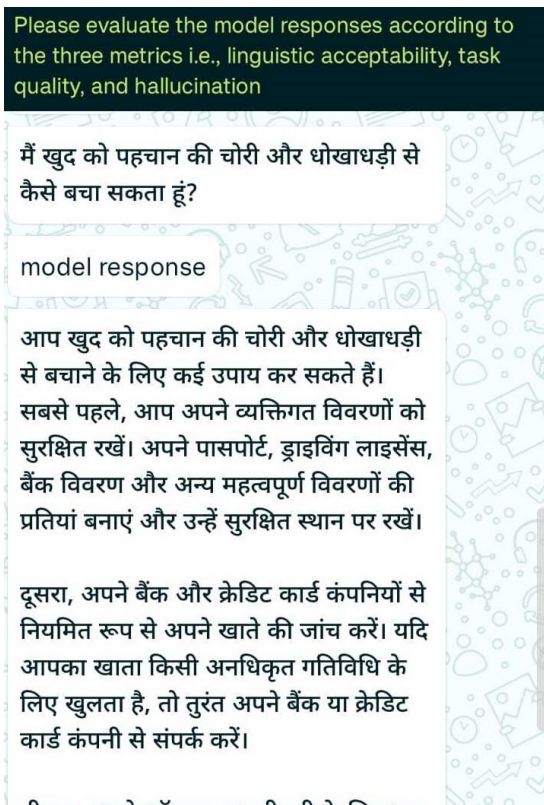


Figure 19: Hindi App screenshots for Direct Assessment human evaluations.

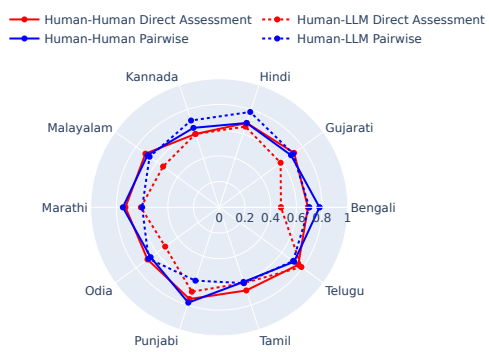


Figure 20: Language-wise PA scores breakdown for Pairwise and Direct Assessment evaluations.