Hindi Reading Comprehension: Do Large Language Models Exhibit Semantic Understanding?

Daisy Monika Lal[1], Paul Rayson[1], Mo El-Haj[1]
[1]School of Computing and Communications, Lancaster University, UK.
Correspondence: d.m.lal@lancaster.ac.uk

Abstract

In this study, we explore the performance of four advanced Generative AI models—GPT-3.5, GPT-4, Llama3, and HindiGPT, for the Hindi reading comprehension task. Using a zero-shot, instruction-based prompting strategy, we assess model responses through a comprehensive triple evaluation framework using the HindiRC dataset. Our framework combines (1) automatic evaluation using ROUGE, BLEU, BLEURT, METEOR, and Cosine Similarity; (2) rating-based assessments focussing on correctness, comprehension depth, and informativeness; and (3) preference-based selection to identify the best responses[1]. Human ratings indicate that GPT-4 outperforms the other LLMs on all parameters, followed by HindiGPT, GPT-3.5, and then Llama3. Preference-based evaluation similarly placed GPT-4 (80%) as the best model, followed by HindiGPT(74%). However, automatic evaluation showed GPT-4 to be the lowest performer on n-gram metrics, yet the best performer on semantic metrics, suggesting it captures deeper meaning and semantic alignment over direct lexical overlap, which aligns with its strong human evaluation scores. This study also highlights that even though the models mostly address literal factual recall questions with high precision, they still face the challenge of specificity and interpretive bias at times.

This is a Hindi reading comprehension task. You will be given a Hindi text passage labeled P and asked questions about it. After reading the passage, respond to each question in Hindi, using complete sentences and the information provided in the passage P.

Example format:
P: [Passage here]
Q: [Question related to passage P, here]
A: [Your complete answer based on passage P]

Figure 1: Instruction-Based Prompting Strategy for Hindi MRC.

1 Introduction

Machine reading comprehension (MRC) in Natural Language Processing (NLP) is the task of making machines retrieve or generate precise and contextually relevant answers from a specific question and a body of text(Chen, 2018; Liu et al., 2019; Baradaran et al., 2022). It has numerous real-world applications, ranging from search engines to educational tools and domain-specific conversational agents or chatbots (Qiu et al., 2019; Baradaran et al., 2022; Kazi et al., 2023). MRC involves understanding the underlying context and is extremely challenging as it requires complex cognitive capabilities like summarising, sequencing, inferencing, and comparing and contrasting facts presented in the given text (Khashabi et al., 2018; Gardner et al., 2019; Sun, 2021). While NLP has seen significant advancements for widely spoken languages, much of the research has left low-resource languages like Hindi underexplored, especially for complex tasks such as MRC. (Jing and Xiong, 2020; Nguyen et al., 2022; Lal et al., 2022).

Hindi, the fourth most-spoken language globally (Yadav, 2023), has witnessed major breakthroughs in NLP technologies in recent years. Nevertheless, as large language models (LLMs) emerge as the cornerstone of NLP research, it is essential to ask: How well do these models understand Hindi? While LLMs perform admirably on surface-level tasks like text generation, text classification, and machine translation (Parida et al., 2024) that do not always require in-depth analysis of comprehension; MRC, that involves nuanced understanding of context, factual information, and reasoning, can serve as a benchmark for assess-

[1]Human annotations available at `https://github.com/dml2611/HindiMRC`.

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 1–10
January 20, 2025.

ing the comprehension abilities of these models for Hindi texts.

In this study, we investigate the performance of four prominent LLMs—GPT-3.5 (Winata et al., 2021), GPT-4 (Ai et al., 2023), HindiGPT[2], and Llama3 (Dubey et al., 2024)—to uncover how well these models perform on Hindi reading comprehension tasks—not just in terms of accurate answers, but the limits of their comprehension and informativeness. To assess the performance of each model, we conducted both automatic and human evaluations (rating-based and preference-based), as shown in Figure 2. The automatic evaluations provide a quantitative assessment of the models, while the human evaluations enable a qualitative assessment of each model's responses. This extensive study allows us to investigate and emphasize where these models thrive and where they fall short, as well as where they need to catch up to human comprehension.

The rest of the paper is organized as follows. Section 2 presents prior Related Work; Section 3 outlines the Methodology; Section 4 states the results, followed by the conclusions and limitations in Sections 5 and 5.

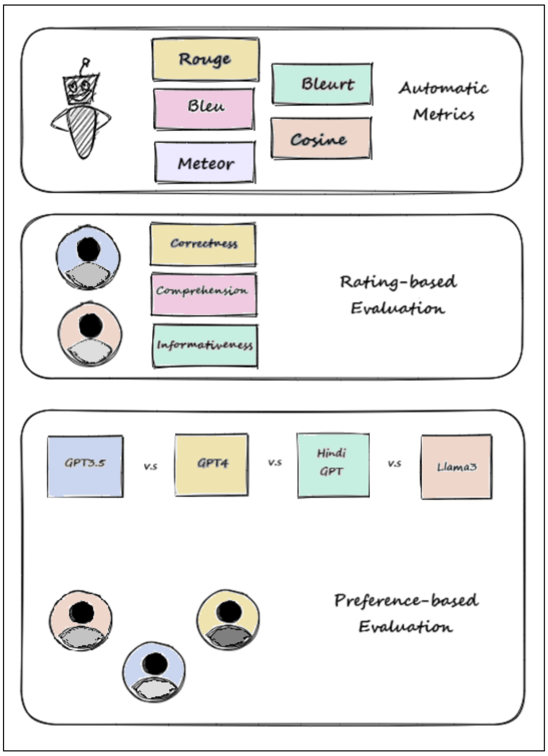

Figure 2: Triple evaluation framework for assessing Hindi reading comprehension in LLMs using automatic and human evaluation methods.

2 Related Work

Researchers in the field of NLP consistently highlight the resource limitations that hinder the development of effective question-answering (QA) systems for low-resource languages such as Hindi (Maddu and Sanapala, 2024; Kumari and Shivhare, 2023; Chaudhari et al., 2024). The scarcity of high-quality, annotated datasets and linguistic tools specifically tailored for Hindi is a significant barrier. State-of-the-art QA models, like BERT and GPT, rely on extensive gold-standard corpora to produce accurate and robust results. However, for Hindi, the availability of such resources remains limited, creating a gap in model performance (Nanda et al., 2016; GUPTA and KHADE, 2020; Khurana et al., 2024). Existing models and datasets are primarily designed for tasks involving short answer spans or multiple-choice responses, which restricts their flexibility.

Another significant challenge is the constrained context length used during model training, primarily due to computational costs associated with handling large amounts of text (Kumar et al., 2022). As a result, the models struggle to grasp the linguistic subtleties of Hindi, such as syntax and morphology, which can reduce overall performance (Ray et al., 2018; Anuranjana, 2021). The complexity of Hindi is further increased by distinct syntactic structures, numerous semantic variants, and prevalent code-mixing (Hindi-English hybrids) in written and spoken forms, add further barriers to QA development (Viswanathan et al., 2019). However, LLMs have shown significant potential in handling diverse languages and can flexibly adapt to code-mixed texts (Brown, 2020; Conneau, 2019; Raffel et al., 2020; Chung et al., 2024), making them potentially valuable tools to address Hindi NLP challenges, like reading comprehension.

3 Methods

In order to examine the comprehension abilities of GPT-3.5, GPT-4, HindiGPT, and Llama3 for Hindi texts, the LLMs were directed to complete the Hindi MRC task using instruction-based prompting on the HindiRC[3] dataset, as described in Sections 3.1 and 3.2, respectively. We evaluated the

[2] *HindiGPT* available at `https://chatgpt.com/g/g-oKGVbNtmC-hindi-gpt`

[3] *HindiRC* available at `https://github.com/erzaliator/HindiRC-Data`

Table 1: LLM responses to Hindi reading comprehension highlighting the challenge of Literal vs. Interpretive Bias.

Passage: जब वह बहुत छोटा था, तभी उसने संयोग(coincidence) से किले का गुप्त(hidden) द्वार(entrance) खोज(found) लिया था। लेकिन न उसने किसी को बताया था और न ही इसके बारे में किसी और को पता था। बस, वह जब चाहे गुप्त(hidden) मार्ग(path) से किले में जा पहुँचता और खेल–कूदकर उसी रास्ते से वापस आ जाता। इस गुप्त(hidden) मार्ग(path) की जानकारी के बिना इस किले पर चढ़ाई करना अत्यंत कठिन था। बालू समझ गया था कि शिवाजी ने इसीलिए इसे अपना गढ़ बनाया, क्योंकि वह इस तरह बना हुआ था कि कोई भी शत्रु इसपर अचानक आक्रमण नहीं कर सकता था। When he was very young, he accidentally discovered the secret entrance of the fort. However, he didn't tell anyone, and no one else knew about it either. So, whenever he wanted, he would go to the fort through the secret route, play around, and return the same way. Without knowledge of this secret path, climbing this fort was extremely difficult. Balu understood that Shivaji had chosen it as his stronghold for this reason, as it was built in such a way that no enemy could launch a sudden attack on it.		
Question	बालू ने क्या रहस्य जान लिया था ? What secret did Balu know??	
GPT-3.5	बालू ने रहस्य जान लिया था कि शिवाजी ने इस किले को इस तरह बनाया था कि कोई भी शत्रु इसपर अचानक आक्रमण नहीं कर सकता था। Balu discovered the secret that Shivaji had built this fort in such a way that no enemy could launch a sudden attack on it.	$\mathcal{C}_o = 1.5$ $\mathcal{C}_h = 1.5$ $\mathcal{I}_f = 1.5$
GPT-4	बालू ने यह रहस्य जान लिया था कि शिवाजी ने किले को अपना गढ़ इसलिए बनाया था क्योंकि इसकी संरचना ऐसी थी कि कोई भी शत्रु इसपर अचानक आक्रमण नहीं कर सकता था। Balu discovered the secret that Shivaji had made the fort his stronghold because its structure was such that no enemy could launch a sudden attack on it.	$\mathcal{C}_o = 1.5$ $\mathcal{C}_h = 1.5$ $\mathcal{I}_f = 1.5$
HindiGPT	बालू ने किले का गुप्त द्वार खोजने का रहस्य जान लिया था। Balu discovered the secret of finding the fort's secret entrance.	$\mathcal{C}_o = 4.0$ $\mathcal{C}_h = 4.0$ $\mathcal{I}_f = 4.0$
Llama3	बालू ने किले का गुप्त द्वार खोज लिया था। Balu discovered the secret entrance of the fort.	$\mathcal{C}_o = 5.0$ $\mathcal{C}_h = 5.0$ $\mathcal{I}_f = 5.0$

performances of the LLMs using both automatic and human assessments, as described in Section 3.3.

3.1 Prompting Strategy

We adopted the zero-shot, instruction-based prompting strategy to evaluate the generalization abilities of the LLMs on the Hindi MRC challenge in a zero-shot setting. All four LLMs were given the same prompt, as shown in Figure 1. This strategy taps into the pre-trained knowledge of the LLM without requiring any task-specific fine-tuning. The structured guidance through explicit instructions directs the LLM on how to carry out a task, even if it hasn't seen it before (Schick and Schütze, 2021; Liu et al., 2023; Chen et al., 2024). Moreover, the instruction template (Passage P, Question Q, Answer A) helps to standardize responses across all LLMs (see Tables 1, 6), enabling direct comparison of performance.

3.2 Dataset

The HindiRC dataset (Anuranjana et al., 2019) is a collection of 24 Hindi reading comprehension passages assembled from two educational websites, Sandeep Barouli[4] and 2classnotes[5]. It comprises 127 questions with corresponding single-sentence answers, manually selected from the passage by the annotator.

[4] *Sandeep Barouli* available at https://sandeepbarouli.com/

[5] *2classnotes* available at https://www.2classnotes.com/

Table 2: The Rating Scale for Human Evaluation. This rating scale grades LLM responses on three criteria: correctness, comprehension depth, and informativeness, with grades ranging from 1 to 5.

Correctness (Factual and Logical Accuracy)	
5 - Entirely correct	no factual errors or inconsistencies.
4 - Mostly correct	minor inaccuracies that don't significantly affect meaning.
3 - Partially correct	contains few inaccuracies that slightly affect meaning.
2 - Mostly incorrect	significant factual or logical errors that compromise accuracy.
1 - Incorrect	fails to address the question with any factual or logical accuracy.
Comprehension (Depth of Understanding)	
5 - Deep understanding	captures nuances and underlying meanings.
4 - Good understanding	covers key concepts though minor details may be missed.
3 - Basic understanding	general answer, missing some deeper context or meaning.
2 - Limited understanding	simplistic or surface-level answer, with key misinterpretations.
1 - No understanding	fails to grasp the main idea or gives an irrelevant answer.
Informativeness (Coverage of Essential Points)	
5 - Fully informative	includes all essential points and relevant details.
4 - Mostly informative	covers most key points, with minor oversights.
3 - Moderately informative	includes some key points but misses several important details.
2 - Minimally informative	misses many important details.
1 - Not informative	fails to include any essential points or details.

Table 3: This table illustrates the scores for automatic evaluation metrics, ROUGE, BLEU, BLEURT, METEOR, and Cosine Similarity (CoS). Here, R1 F1, R2 F1, and RL F1 refer to ROUGE-1 F1, ROUGE-2 F1, and ROUGE-L F1 Scores, respectively.

	Metric	GPT-3.5	GPT-4	HindiGPT	Llama3
n-gram matching	**R1 F1**	0.540	0.512	0.540	0.533
	R2 F1	0.405	0.401	0.404	0.433
	RL F1	0.510	0.494	0.515	0.516
	BLEU	0.348	0.317	0.358	0.373
semantic similarity	**BLEURT**	0.530	0.431	0.497	0.458
	METEOR	0.515	0.516	0.507	0.508
	CoS	0.922	0.924	0.924	0.914

3.3 Evaluation Strategy

The evaluation setup includes seven automatic metrics and three human evaluation rating scales. We also use preference-based human evaluation to gain additional insights into human preferences.

3.3.1 Automatic Assessment

The automatic assessment was carried out using five different metrics: 1) ROUGE (Lin, 2004) predominantly assesses recall by calculating overlapping n-grams (ROUGE-1), word pairs (ROUGE-2), and word sequences (ROUGE-L), between machine-generated and reference responses. 2) BLEU (Papineni et al., 2002) compares the n-grams in the machine-generated response to those in the reference response. Typically used in translation, but can also assess how effectively a machine-generated response captures the key terms. 3) BLEURT (Sellam et al., 2020) is a learned metric that addresses the shortcomings of conventional n-gram-based metrics like BLEU and ROUGE. It leverages a pre-trained transformer model to determine the semantic similarity between machine-generated and reference responses. 4) METEOR (Banerjee and Lavie, 2005) measures semantic similarity using synonyms, stemming, and partial matches, and has a high correlation with human judgment. 5) Cosine Similarity (CoS) (Rahutomo et al., 2012) compares model-generated responses to reference answers using word embeddings, judging similarity in sense rather than precise word

Table 4: Preference-based selection results for three annotators $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$.

	GPT-3.5	GPT-4	HindiGPT	Llama3
$\mathcal{H}_1$	75%	75%	73%	65%
$\mathcal{H}_2$	73%	83%	75%	68%
$\mathcal{H}_3$	68%	83%	73%	68%
Avg	**72%**	**80%**	**74%**	**67%**

match. We employed FastText Hindi[6] embeddings to compute CoS.

3.3.2 Human Evaluation

Two human evaluators rated responses based on correctness, comprehension depth, and informativeness. Another set of three evaluators determined the best responses based on overall preferences. All evaluations were conducted on a randomly selected set of 40 questions from eight distinct passages.

a) Rating-based Evaluation or (Likert-rating) involves grading each response individually based on predefined criteria, such as correctness, comprehension, and informativeness (described in Table 2). This strategy allows evaluators to express the extent to which each criterion is met. Correctness ensures factual and logical accuracy, which is fundamental to comprehension quality. Comprehension measures the depth of understanding, indicating whether the LLM genuinely understands the underlying context rather than providing shallow responses. Informativeness evaluates the information coverage, ensuring that important facts and nuances are not overlooked.

b) Preference-based Selection (or Best-Answer Selection) This approach requires assessors to select the answers they find most satisfactory among the provided options. This method offers a more precise indication of which models consistently generate higher-quality responses, allowing for a direct assessment of performance based on the overall quality of response.

4 Results

The overall results of human and automatic evaluations, along with the inter-rater reliability, are covered in Sections 4.1, 4.2, and 4.3, respectively.

4.1 Automatic Assessment

The results of the automatic assessment (Table 3) demonstrate that GPT-3.5 (BLEURT = 0.530, CoS = 0.922) and GPT-4 (BLEURT = 0.431, CoS = 0.924) score better on semantic metrics, suggesting that they prioritize meaning over exact wording and structure. This indicates that for tasks seeking nuanced interpretation and linguistic mobility, these LLMs might be a preferable choice. HindiGPT consistently performs well across ROUGE (R1 F1 = 0.540, R2 F1 = 0.404, and RL F1 = 0.515), BLEU (0.358), and cosine similarity (0.924), demonstrating that it successfully captures meaning. This makes it suitable for tasks where semantic comprehension is essential. Llama3 exhibits notable word sequence and phrase-matching abilities, which could signify higher proficiency and coherence at the phrase-level. It also scores well in BLEU (0.373) and ROUGE metrics (R1 F1 = 0.533, R2 F1 = 0.433, and RL F1 = 0.516), suggesting that tasks where exact match is preferred to subtle understanding may be its ideal fit.

4.2 Human Assessment

The rating-based evaluation sheds light on how well each LLM performed for each metric, based on both annotators' ratings and the confidence intervals (CIs) around these ratings (see Table 5).

Correctness ($\mathcal{C}_o$): GPT-4 ($\mathcal{A}_1 = 4.725 \pm 0.029$ and $\mathcal{A}_2 = 4.700 \pm 0.035$) has the highest mean scores for both annotators, with very narrow CIs, signifying high precision and annotator confidence in ratings. HindiGPT scores ($\mathcal{A}_1 = 4.650 \pm 0.026$ and $\mathcal{A}_2 = 4.600 \pm 0.026$) fall closely behind GPT-4, implying good precision but slightly lower than GPT-4. GPT-3.5 ($\mathcal{A}_1 = 4.625 \pm 0.039$ and $\mathcal{A}_2 = 4.550 \pm 0.038$) and Llama3 ($\mathcal{A}_1 = 4.575 \pm 0.039$ and $\mathcal{A}_2 = 4.550 \pm 0.029$) have comparatively lower mean scores than GPT-4 and HindiGPT. Llama3 exhibited a wider CI, implying greater variation in the perception of

[6] *fasttext-hi-vectors* available at https://huggingface.co/facebook/fasttext-hi-vectors

Table 5: Human evaluation results for LLM performance across metrics (correctness ($\mathcal{C}_o$), comprehension ($\mathcal{C}_h$), and informativeness ($\mathcal{I}_f$)) with Mean Scores and Confidence Intervals for each model, alongside the Cohen's Kappa (κ) statistic for inter-annotator agreement between annotators $\mathcal{A}_1$ and $\mathcal{A}_2$.

		GPT-3.5	**GPT-4**	**HindiGPT**	**Llama3**
$\mathcal{A}_1$	$\mathcal{C}_o$	4.625 ± 0.039	4.725 ± 0.029	4.650 ± 0.026	4.575 ± 0.039
	$\mathcal{C}_h$	4.620 ± 0.039	4.775 ± 0.034	4.650 ± 0.027	4.600 ± 0.039
	$\mathcal{I}_f$	4.525 ± 0.041	4.750 ± 0.028	4.650 ± 0.028	4.550 ± 0.041
$\mathcal{A}_2$	$\mathcal{C}_o$	4.550 ± 0.038	4.700 ± 0.035	4.600 ± 0.026	4.550 ± 0.029
	$\mathcal{C}_h$	4.625 ± 0.036	4.850 ± 0.032	4.675 ± 0.027	4.450 ± 0.036
	$\mathcal{I}_f$	4.575 ± 0.033	4.750 ± 0.034	4.600 ± 0.028	4.450 ± 0.036
κ	$\mathcal{C}_o$	0.634	0.808	0.695	0.520
	$\mathcal{C}_h$	0.508	0.696	0.840	0.675
	$\mathcal{I}_f$	0.709	0.712	0.694	0.682

correctness.

Comprehension ($\mathcal{C}_h$): GPT-4 receives the highest scores for this measure, particularly from $\mathcal{A}_2 = 4.850 \pm 0.032$ ($\mathcal{A}_1 = 4.775 \pm 0.034$), signifying GPT-4's strong comprehension abilities, particularly with a low CI, suggesting annotators found its answers consistently comprehensive. HindiGPT performs well too, scoring $\mathcal{A}_1 = 4.650 \pm 0.027$ and $\mathcal{A}_2 = 4.675 \pm 0.027$, with consistently high comprehension scores, although slightly lower than GPT-4. GPT-3.5 ($\mathcal{A}_1 = 4.620 \pm 0.039$ a and $\mathcal{A}_2 = 4.625 \pm 0.036$) and Llama3 ($\mathcal{A}_1 = 4.600 \pm 0.039$ and $\mathcal{A}_2 = 4.450 \pm 0.036$) yield slightly lower scores. Llama3 has a lower comprehension score, demonstrating some variation in perceived comprehension quality.

Informativeness ($\mathcal{C}_o$): GPT-4 obtains the highest scores ($\mathcal{A}_1 = 4.750 \pm 0.028$ and $\mathcal{A}_2 = 4.750 \pm 0.034$), suggesting strong information coverage in responses. HindiGPT ($\mathcal{A}_1 = 4.650 \pm 0.028$ and $\mathcal{A}_2 = 4.600 \pm 0.028$) follows GPT-4, exhibiting adequate but slightly less information coverage. GPT-3.5 ($\mathcal{A}_1 = 4.525 \pm 0.041$ and $\mathcal{A}_2 = 4.575 \pm 0.033$) has slightly lower scores than HindiGPT, indicating that it may overlook a few crucial details. Llama3 scores the lowest, implying having the least information coverage and some fluctuation in perceived quality.

Preference-based evaluation (see Table 4) revealed that GPT-4 was consistently favoured by the three annotators, with an average score of 80%. Its high preference indicates that, in terms of human judgment, GPT-4's answers were relevant, demonstrating an excellent ability to provide accurate and consistent responses to questions. With an average score of 72%, GPT-3.5 was slightly lower than GPT-4 but still obtained significant preference, indicating that it might have occasionally fallen short of GPT-4. With a 74% average, HindiGPT performed in the competitive range of GPT-4 and around GPT-3.5. Its consistent ranking indicates that it offered replies that were linguistically and semantically appropriate. Llama3 received the lowest preference from the annotators, with an average of 67%.

4.3 Inter-Annotator Agreement

We apply Cohen's Kappa coefficient (κ) to gauge the inter-annotator agreement for rating-based evaluation (McHugh, 2012). We compute κ per metric for all question-answer pairs in the HindiRC evaluation set. Finally, we assess the reliability for each LLM separately to determine agreement per metric between evaluators (see Table 5).

Correctness ($\mathcal{C}_o$): GPT-4 (0.808) had the highest κ for correctness, while Llama3 (0.520) had the lowest. This suggests that GPT-4 responses were more reliable and easier for annotators to agree on. Annotators were relatively in agreement on the correctness of this GPT-3.5 (0.634) and HindiGPT (0.695) responses.

Comprehension ($\mathcal{C}_h$): HindiGPT (0.840) obtained the highest κ for comprehension, suggesting that the responses were generated with in-depth understanding of the context that leads to accurate answers. In contrast, GPT-3.5 (0.508)

Table 6: LLM responses to Hindi reading comprehension highlighting the challenge of fact recall and specificity in responses.

<table>
<tr><td colspan="3">Passage: अमीर खुसरो न केवल साधारण जन के लिए कविताएँ लिखते थे, बल्कि उस समय के बादशाहों की प्रशंसा में भी शायरी लिखते थे। वे युद्ध में भी भाग लेते थे और रण क्षेत्र में अपनी बहादुरी का जौहर भी दिखलाते थे। उनकी कविता और वीरता से प्रभावित होकर अनेक बादशाहों ने उन्हें अपने दरबार में रखा। इस प्रकार प्रजा से लेकर राजा तक अर्थात्झों पड़ी से लेकर महल तक अमीर खुसरो की प्रसिद्धि फैल गई। उनकी प्रसिद्धि का एक कारण उनका संगीतकार होना था। वे संगीत के ज्ञाता(expert in music) थे। उन्हें राग–रागिनियों का अच्छा ज्ञान(knowledge) था और उन्होंने कई राग–रागिनियों की रचना भी की।
Amir Khusro not only wrote poems for the common people but also composed poetry in praise of the kings of his time. He participated in wars and displayed his bravery on the battlefield as well. Impressed by his poetry and valor, many kings kept him in their courts. Thus, Amir Khusro's fame spread from the common folk to royalty, from humble huts to grand palaces. One reason for his fame was his talent as a musician. He was a knowledgeable musician. He had a good knowledge of Raga-Raginis. He also composed many Raga-Raginis.</td></tr>
<tr><td>Question</td><td colspan="2">अमीर खुसरो किसके ज्ञानी माने जाते थे ?
Amir Khusro was considered knowledgeable in what?</td></tr>
<tr><td>GPT-3.5</td><td>अमीर खुसरो संगीत के ज्ञाता माने जाते थे।
Amir Khusro was considered knowledgeable in music.</td><td>$\mathcal{C}_o = 5.0$
$\mathcal{C}_h = 4.5$
$\mathcal{I}_f = 4.0$</td></tr>
<tr><td>GPT-4</td><td>अमीर खुसरो संगीत के ज्ञानी माने जाते थे, उन्हें राग–रागिनियों का अच्छा ज्ञान था और उन्होंने कई राग–रागिनियों की रचना भी की।
Amir Khusro was considered knowledgeable in music; he had a deep understanding of Raga-Raginis and also composed many Raga-Raginis.</td><td>$\mathcal{C}_o = 5.0$
$\mathcal{C}_h = 5.0$
$\mathcal{I}_f = 5.0$</td></tr>
<tr><td>HindiGPT</td><td>अमीर खुसरो राग–रागिनियों के ज्ञानी माने जाते थे।
Amir Khusro was considered knowledgeable in Raga-Raginis.</td><td>$\mathcal{C}_o = 3.0$
$\mathcal{C}_h = 3.0$
$\mathcal{I}_f = 3.0$</td></tr>
<tr><td>Llama3</td><td>अमीर खुसरो संगीत के ज्ञानी माने जाते थे, उन्हें राग–रागिनियों का अच्छा ज्ञान था।
Amir Khusro was considered knowledgeable in music; he had a good understanding of Raga-Raginis.</td><td>$\mathcal{C}_o = 5.0$
$\mathcal{C}_h = 5.0$
$\mathcal{I}_f = 5.0$</td></tr>
</table>

showed moderate agreement, indicating a degree of variability in perceived comprehension quality. GPT-4 (0.696) and Llama3 (0.675) showed substantial agreement, indicating that annotators were generally aligned.

Informativeness ($\mathcal{I}_f$): All models exhibited relatively similar kappa scores for informativeness, with substantial agreement across GPT-3.5 (0.709), GPT-4 (0.712), HindiGPT (0.694), and Llama3 (0.682). This consistency indicates that informativeness was relatively straightforward to assess, resulting in consistent alignment amongst annotators.

5 Conclusion

In this study, we use a novel triple assessment framework to compare the performance profiles of LLMs, GPT-3.5, GPT-4, Llama3, and HindiGPT for Hindi reading comprehension. The ability of GPT-4 to generate contextually relevant and meaningful responses is demonstrated by its preference rating of 80%, which consistently outperformed competing models across all human-evaluated metrics—correctness, comprehension depth, and informativeness. With competitive scores, particularly in correctness and comprehension, HindiGPT and GPT-3.5 trailed closely behind. GPT-3.5 was somewhat preferred above HindiGPT for perceived understanding and precise responses.

The results of automatic evaluations presented a contrasting picture, indicating fewer exact matches with reference texts, particularly for GPT-4 with lower n-gram metric scores (ROUGE and BLEU). The high human evaluation scores of GPT-4 are consistent with its superior alignment with the underlying meaning as measured by semantic metrics (BLEURT and Cosine Similarity), which show a greater grasp of the text than surface-level similarity. This comparison of automatic and human evaluations highlights the significance of semantic-based metrics and human evaluations for precisely assessing the level of a LLM's comprehension, particularly in non-English languages like Hindi.

Limitations

Our research reveals limitations in some of the metrics which do not align well with human assessment. As well as limitations of the domains or topics expressed within the dataset, our results are tied to the current versions of the four specific LLMs that we have used in our experiments. In future work, we will test other open-source models. In some cases, we find that models have a tendency to overinform in their answers, and we will investigate further techniques to reduce this.

Literal vs. Interpretive Responses: In Table 1, the question "*What secret did Balu discover?*" seeks a factual answer about the "गुप्त द्वार""(*secret entrance*)." HindiGPT and Llama3 are more literal in answering the question, providing answers that adher to exact phrases from the passage. However, GPT-3.5 and GPT-4 misinterpret the question's focus and provide an interpretative response about the strategic purpose of the fort's design, showing an interpretive bias (Sheng et al., 2019; Bender et al., 2021). This disparity between question focus and model response arises because of the models' tendency to prioritize interpretations and contextual meaning over literal facts. The likelihood of LLMs adding extraneous information is a common issue with models in open-ended tasks (Koul, 2023).

Fact Recall and Specificity in Responses: In Table 6, in response to the question "*Amir Khusro was considered knowledgeable in what?*", the factual answer is संगीत(*music*), as the passage makes it abundantly evident that Khusro was an expert in music and that his knowledge of Raga-Raginis was a core reason for his fame. Yet, the models provided responses with varying degrees of specificity highlighting a gap in fact recall (Petroni et al., 2019). GPT-3.5 states that Amir Khusro was knowledgeable in music, but it omits the details of Raga-Raginis, giving a less comprehensive response. GPT-4 correctly mentions Amir Khusro's expertise of music and Raga-Raginis, and it also adds that he composed many of them. HindiGPT generates a partially correct response. It focuses on "Raga-Raginis" but omits the broader aspect of Amir Khusro's music knowledge, missing the broader context of his musical knowledge and his composition of them. Is informative but lacks the extra detail about his compositions. Llama3's provides a good amount of detail, mentioning both music and Raga-Raginis, but omits the fact that Amir Khusro composed them.

References

Lets Build Your Ai, Corporate Ai, and Restaurant Ai. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kaveri Anuranjana. 2021. *Towards building Question Answering Resources for Hindi*. Ph.D. thesis, International Institute of Information Technology Hyderabad.

Kaveri Anuranjana, Vijjini Rao, and Radhika Mamidi. 2019. Hindirc: a dataset for reading comprehension in hindi. In *0th International Conference on Computational Linguistics and Intelligent Text*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?□□. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Deepti A Chaudhari, Rahul Shrivastava, and Sanjeevkumar Angadi. 2024. A survey on conversational ai question answering system for low resource language. *Journal of Electrical Systems*, 20(6s):2531–2540.

Danqi Chen. 2018. *Neural reading comprehension and beyond*. Stanford University.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112.

SOMIL GUPTA and NILESH KHADE. 2020. Bert based multilingual machine comprehension in english and hindi. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, 19(1).

Yimin Jing and Deyi Xiong. 2020. Effective strategies for low-resource reading comprehension. In *2020 International Conference on Asian Language Processing (IALP)*, pages 153–157. IEEE.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2023. A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, 56(Suppl 2):2509–2569.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Khushboo Khurana, Rachita Bharambe, Hardik Dharmik, Krishna Rathi, and Mayur Rawte. 2024. A textual question answering and handwritten answer evaluation system for hindi language. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 28(3):435–455.

Nimrita Koul. 2023. *Prompt Engineering for Large Language Models*. Nimrita Koul.

Shailender Kumar et al. 2022. Bert-based models' impact on machine reading comprehension in hindi and tamil. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 1458–1462. IEEE.

Pooja Kumari and Rakesh Shivhare. 2023. Study of various approaches used for machine reading comprehension in question answering systems. *International Journal of Technology Research and Management*.

Bechoo Lal, G Shivakanth, Arun Bhaskar, M Bhaskar, Ashish, and Deepak Kumar Panda. 2022. Critical review on machine reading comprehension (mrc) developments: From high resource to low resource languages. In *International Advanced Computing Conference*, pages 341–352. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Sandeep Maddu and Viziananda Row Sanapala. 2024. A survey on nlp tasks, resources and techniques for low-resource telugu-english code-mixed text. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Garima Nanda, Mohit Dua, and Krishma Singla. 2016. A hindi question answering system using machine learning approach. In *2016 international conference on computational techniques in information and communication technologies (ICCTICT)*, pages 311–314. IEEE.

Bach Hoang Tien Nguyen, Dung Manh Nguyen, and Trang Thi Thu Nguyen. 2022. Machine reading comprehension model for low-resource languages and experimenting on vietnamese. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 370–381. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shantipriya Parida, Shakshi Panwar, Kusum Lata, Sanskruti Mishra, and Sambit Sekhar. 2024. Building pre-train llm dataset for the indic languages: a case study on hindi. *arXiv preprint arXiv:2407.09855*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A survey on neural machine reading comprehension. *arXiv e-prints*, pages arXiv–1906.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1. University of Seoul South Korea.

Santosh Kumar Ray, Amir Ahmad, and Khaled Shaalan. 2018. A review of the state of the art in hindi question answering systems. *Intelligent Natural Language Processing: Trends and Applications*, pages 265–292.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.

Kai Sun. 2021. *Machine reading comprehension: challenges and approaches*. Cornell University.

Sujith Viswanathan, M Anand Kumar, and KP Soman. 2019. A comparative analysis of machine comprehension using deep learning models in code-mixed hindi language. *Recent Advances in Computational Intelligence*, pages 315–339.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Vinod Kumar Yadav. 2023. Impact of globalization on english and hindi languages: An analysis. *Anand Bihari*, page 230.