Does Reasoning Kill the Joke? Long-Context Humor Understanding in Hindi

Kaveri Anuranjana*[1] Navya Shrivastava*[1] Atharv Johar*[1] Rishabh Sabharwal*[2,5]
Gautam Ranka[3] Aryan Lunawat[1,4] Punit Rathore[5] Radhika Mamidi[1]

[1]IIIT Hyderabad [2]University of Edinburgh [3]VNIT Nagpur
[4]JK Lakshmipat University [5]IISc Bangalore

{kaveri.a, navya.shrivastava, atharv.johar}@research.iiit.ac.in

Abstract

Verbal humor involves reasoning through complex conversational contexts. Although LLMs have achieved strong performance on English humor datasets, their ability to interpret humor in Hindi remains unexplored. In this paper, we evaluate Hindi humor for which we extract dialogues from humorous video clips. We use a pipeline that transforms video content into detailed textual streams, including dialogue transcripts and scene descriptions, allowing reasoning over inputs exceeding 2,000 words. We test various LLMs, from efficient edge models (Qwen-2.5-7B, Qwen-3-7B, Gemma-3-27B) to Indic-focused models (Sarvam-M-24B) and large frontier models (Llama-3.1-70B, Gemini-2.0-Flash). Our findings show a concave performance pattern in long context understanding, with reasoning quality peaking at moderate lengths (250–750 words) and declining at higher context lengths. While increasing model size generally improves performance, smaller LLMs face instructional and linguistic issues. Adding a diversity metric can capture hallucinatory failures. Furthermore, smaller, Hindi-focused models can compete with much larger generalist models. Importantly, our evaluation reveals that conversational humor is a challenge for even specialized models, making HinS a valuable benchmark for advancing research in *Hindi Long-Context Humor Reasoning*.

1 Introduction

As a high-level pragmatic task, verbal humor understanding (Mihalcea and Strapparava, 2006; Alexander, 1997) is grounded in contextual cultural common ground, yet its evaluation is restricted to Western, high-resource languages (Hasan et al., 2019; Hyun et al., 2024). This Anglocentric dataset bias creates a blind spot in assessing **linguistic equity**, a model's capability to maintain performance across diverse societies (Saha et al., 2025), leaving it unclear if pragmatic reasoning extends to conversational Hindi beyond the high-resource English envelope. Although recent efforts have expanded resources for Indic NLP (Doddapaneni et al., 2023; Kakwani et al., 2020), Hindi remains severely under-served in multimodal humor datasets. Meanwhile existing approaches in Indian humor are predominantly oriented towards simpler objectives such as detection (Chauhan et al., 2021; Bedi et al., 2021).

We address this gap with *HinS* [1], a novel dataset to evaluate reasoning on Hindi long-context humor sourced from native conversations. These conversations sourced from sitcoms and stand-up comedy lend to conversational humor (Dynel, 2011). As a dialogue-centric benchmark, we adopt a *textual approach* inspired by (Hyun et al., 2024). Video data is converted in textual (audio, visual and transcript) inputs, allowing text-LLMs to be tested on their pragmatic reasoning in long, descriptive conversations.

A comparative evaluation of English versus Hindi humor conducted on HinS reveals a glaring **understandability gap** across all models. While recent works (Hyun et al., 2024; Hasan et al., 2019) may lead us to believe that LLMs can understand humor reasonably well, when trying to evaluate it in a cultural setup, we discover quite the opposite. Standard surface and semantic-level similarity metrics misleadingly overstate a model's actual pragmatic competence. Furthermore, *hallucinatory failures* can be effectively identified by *combining a linguistic diversity metric with standard quality metrics*. **Hindi conversational humor understanding** still remains an **open challenge** even for language specialized models.

Our detailed analysis, baselining models across scale on HinS reveals their limits in understanding Hindi humor, particularly in lengthier contexts.

[1]Dataset that has been anonymized: https://osf.io/hsdx8/overview?view_only=dc86594301e844ebb78221ca5887f81c

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 177–186

The detailed input structure of HinS in the form of visual and audio cues coupled with fairly long conversations, make the task a **long-context understanding challenge**. We identify a downward concave trend where interestingly, models struggle with long context humor and short context humor. Lastly, we test the effect of scale and language-specialization and find that humor reasoning. Our findings demonstrate that humor reasoning benefits from increased model scale; moreover a multilingual, Indic model demonstrates better reasoning on our Hindi dataset.

2 Related Work

2.1 Visual Humor Understanding

Early work in visual humor focused on simple classification tasks, such as identifying whether an image or video clip is funny (Hasan et al., 2019; Castro et al., 2019). Recently, the field has shifted towards *explanation*, with datasets like SMILE (Hyun et al., 2024) challenging models to produce natural language descriptions of the reasons behind laughter. Other explanation-oriented tasks have explored humor in static contexts, such as the New Yorker caption contest (Hessel et al., 2023). The field has transitioned from traditional rule-based methods to deep learning-based fusion architectures, as detailed in the survey (Joshi et al., 2017). Recent advances emphasize the integration of large-scale pretrained transformers to capture higher-order semantics (Poria et al., 2017). Our work aims to extend this by moving from a "detection" to a "reasoning" framework that challenges the cognitive depth of contemporary LLMs. Moreover, the datasets remain heavily biased towards English-speaking cultures, often failing to capture the pragmatic nuances of non-Western contexts.

2.2 Hindi Humor and Meme models

Research on Hindi humor mainly focuses on sarcasm detection in code-mixed social media text (Swami et al., 2018; Vijay et al., 2018) and the classification of multimodal memes (Patwa et al., 2023; Maity et al., 2022). These efforts highlight the linguistic challenges of Hinglish, such as non-standard grammar and script-mixing. While recent multimodal datasets like M2H2 (Chauhan et al., 2021) and MaSaC (Bedi et al., 2021) have introduced conversational humor and sarcasm detection in Hindi sitcoms, they remain centered on binary classification ("is-funny"). As noted in reviews (Sidhu et al., 2023; Patwa et al., 2020), sentiment analysis has advanced in Hindi, but there remains a significant lack of benchmarks that explore why Hindi video content is humorous. Our work addresses this gap by focusing on reasoning and *natural language explanations for laughter in native video Hindi conversations*.

3 The HinS Dataset

To evaluate the models, we introduce the *HinS* benchmark which includes a curated test set of 1,046 instances (Table 1) sourced from Indian media, specifically sitcoms and stand-up comedy video clips. These domains were selected to ensure long-context Hindi dialogues. Of these, 676 instances (approximately 64.5%) are distributed under Creative Commons (CC) licenses, which allow redistribution and reuse. The remaining 370 clips (approximately 35.4%) are under standard copyright restrictions. Unlike prior humor datasets such as SMILE (Hyun et al., 2024) and UR-FUNNY (Hasan et al., 2019), which predominantly rely on copyrighted material, HinS is the first humor dataset in which the majority of clips are CC-licensed, standing in contrast to the prevailing trend of curating datasets from commercial media. Importantly, the CC-licensed subset alone constitutes a substantial and self-contained benchmark, preserving the full diversity of discourse types (i.e., stand-ups, sitcoms and movies), ensuring that reproducible evaluation remains viable even in the absence of the restricted material.

Total No. of Videos	44
Total No. of Clips	1046
No. of Video Segments	9531
Avg. Clips per Video	23.7
Avg. Segments per Clip	9.11

Table 1: Clip-based and word distribution statistics of HinS

As we focus on text-based LLM evaluation, we convert the video data into textual streams to capture visual and audio modalities (overview of all steps and reasoning pipeline is in Figure 1):

- `Step 1.` **Transcripts:** Time-aligned transcripts that capture dialogue (that consist of code-mixed Hindi-English (Hinglish)), are extracted from clips.
- `Step 2.` **Scene Descriptions (SD):** To provide context usually found in the visual modal-

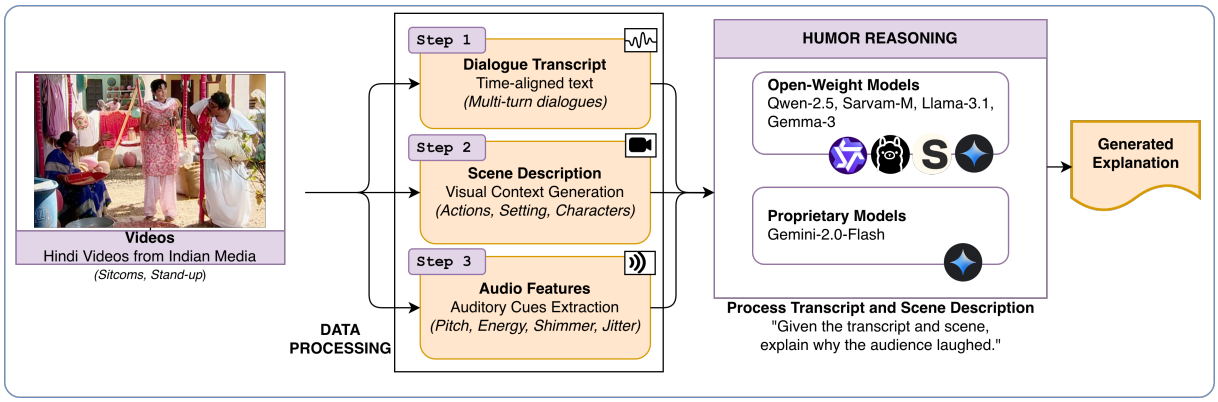

Figure 1: **Dataset Creation and Benchmarking Inference Outline.** *(i)* To evaluate humor reasoning without the computational overhead of video processing, *Transcripts*/dialogues are extracted from video clips. This is followed by extracting *Scene Descriptions* (capturing visual context) using Frontier captioning models. Finally, Audio features are extracted based on (Hyun et al., 2024)'s implementation. *(ii)* For baselining long-context humor reasoning these inputs are fed into a suite of *Open-Weight* and *Proprietary* LLMs to benchmark their ability to generate Hindi humor reasoning (Section 3).

ity, we generate dense textual descriptions of a scene (time aligned with the transcripts) using a Frontier captioning model.

- `Step 3`. **Audio Features:** Similar to (Hyun et al., 2024), audio features are extracted for each segment of each clip.

Following (Hyun et al., 2024), we frame *Explanation Generation* by maximizing $P(Y|X)$, where Y explains the laughter given input $X = T \oplus S$ (transcript concatenated with scene description). This generation objective is an improvement over binary classification, utilizing human-verified generated ground truth (Y_{GT}). This collection of automated ground truth values was verified by native Hindi-speaking university students. From the randomly selected 200 ground truth samples, 86.65% were found to be correct (with inter-annotator agreement (IAA) of 0.68).

4 Model Baselines

To evaluate the landscape of current capabilities in Hindi humor understanding, we benchmark a diverse set of LLMs. Baselines are categorized into two paradigms: *Open-Weight Models* which offer transparency and reproducibility; and *Proprietary Models* which represent the current SOTA in commercial APIs. This distinction allows assessment of accessible, small LLMs against massive closed-source systems.

4.1 Open-Weight Models

We select four high-performing open-weight models that span a wide range of parameter scales and training methodologies.

Qwen-2.5-7B (Team et al., 2024): Qwen-2.5, known for its efficient and lightweight architecture, acts as our baseline for efficiency. Although it is compact at 7B, it shows robust multilingual capabilities on standard benchmarks. We use this model to evaluate if smaller, consumer-grade hardware can handle complex pragmatic tasks given ample textual context.

Qwen-3-7B (Team, 2025): Qwen-3-7B is the latest 7-billion-parameter release in Alibaba's Qwen series, succeeding Qwen-2.5-7B. Its defining characteristic is a hybrid reasoning architecture supporting both a deliberative "thinking" mode for chain-of-thought inference and a fast "non-thinking" mode for direct responses. Despite its compact scale, it demonstrates competitive multilingual performance and strong instruction-following.

Sarvam-M-24B (AI et al., 2024): Sarvam-M is a model with 24 billion parameters, specially pre-trained on a large corpus of Indic languages. Unlike typical generalist models, in which Hindi is only a small part of the training data, Sarvam-M enables us to assess the effect of *targeted domain adaptation*. By comparing it with generalist models, we explore whether high-quality, language-specific pre-training can offset the need for extremely large-scale data.

Gemma-3-27B (Team et al., 2025b): Gemma-3, positioned in the mid-scale category, features

Model	BLEU	METEOR	ROUGE (1/2/L)	BERTScore	TTR
Open-Weight Models					
Qwen-2.5-7B	0.0399	0.1899	0.25 / 0.09 / 0.21	0.7828	0.1116
Qwen-3-7B	0.0410	0.2163	0.26 / 0.09 / 0.22	0.7994	0.1019
Sarvam-M-24B	0.0471	0.2274	0.26 / 0.08 / 0.21	0.7890	0.0961
Gemma-3-27B	0.0380	0.2220	0.30 / 0.08 / 0.24	0.8019	0.1377
Llama-3.1-70B	0.0523	0.2334	0.30 / 0.10 / 0.24	0.7910	0.0971
Proprietary Models					
Gemini-2.0-Flash	**0.0566**	**0.2869**	**0.31 / 0.10 / 0.24**	**0.8075**	0.1226

Table 2: Benchmarking LLMs on Hindi Humor Reasoning. Models are provided with transcripts and scene descriptions. Rows highlighted in light blue indicate model categories.

modern architecture that balances computational efficiency with reasoning depth. As an open model derived from the same research as the Gemini family, it provides an important benchmark for examining how effectively "distilled" reasoning abilities transfer to non-English cultural settings in comparison to larger teacher models.

Llama-3.1-70B (Dubey et al., 2024): Llama-3.1-70B is widely considered the current standard for open-weight reasoning and serves as our primary benchmark for "scale". Its extensive pre-training on diverse global data enables us to test whether large parameter counts are essential for capturing the "long tail" of cultural common sense needed to explain humor.

4.2 Proprietary Models

Gemini-2.0-Flash (Team et al., 2025a): We include Gemini-2.0-Flash to represent the upper limit of current commercial capabilities. As a frontier model designed with native multimodal reasoning (here implemented via textual input), it benefits from advanced instruction-following and large context windows. Benchmarking against Gemini allows us to gauge the performance gap between the best available open models and proprietary commercial systems in the area of low-resource language understanding.

5 Evaluation Metrics

Model performance is assessed by measuring the overlap between generated explanations and Ground Truth (GT) explanations. We report BLEU, ROUGE-1/2/L, and METEOR to quantify surface-level similarity. Additionally, we use BERTScore to measure semantic alignment between generated explanations and GT.

6 Humor reasoning is more challenging in Hindi

Humor understanding, is a socio-linguistic task grounded in cultural context. Therefore, it can be a probe for linguistic equity i.e., the capability of a model to maintain performance across diverse cultures (Saha et al., 2025). For a comparison between English and Hindi humor reasoning, we specifically choose models with either strong multilingual (Gemma-3-27B) or thinking (Deepseek-R1-52B) capabilities or both (Gemini-2.0-Flash). We extract textual, audio and visual (SD) features and reason over these combined inputs to obtain humor reasoning. A comparison of 150 random samples between the English-centric SMILE and Hindi *HinS* reveals a constant understandability gap (Figure 2 and Figure 3) for all models. The equity gap is largest for DeepSeek-R1-0528, which exhibits a 38.3% decline in BLEU (0.1138 $\rightarrow$ 0.0702) and a significant drop in BERTScore (0.8239 $\rightarrow$ 0.7958), indicating that high general-purpose reasoning capacity does not translate to humor reasoning in non-Western contexts. Gemma-3-27B's degradation of BLEU score (0.0711 $\rightarrow$ 0.0436) indicates that even mid-scale multilingual models fail to preserve linguistic equity when faced with the long-context Hindi dialogue. Gemini-2.0-Flash demonstrates the highest equitability (however, still lacks Hindi reasoning) due to superior reasoning capabilities and training which focuses on Indic languages more than other frontier, commercial models. This demonstrates that Hindi conversational reasoning is a two-fold challenge that requires both *multilingual* and *reasoning* based training.

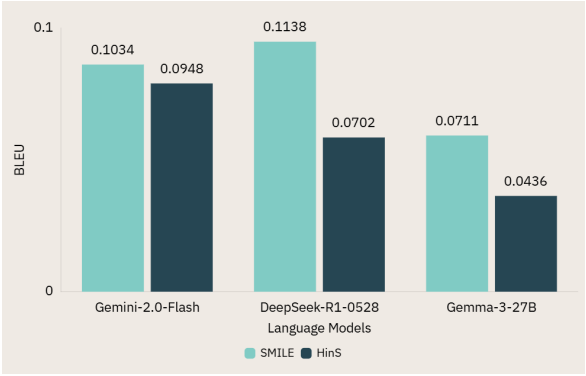

Figure 2: BLEU scores English and Hindi humor reasoning (SMILE vs. HinS). Consistently low performance of all models demonstrates the *equity-gap* of LLMs in Hindi cultural understanding.

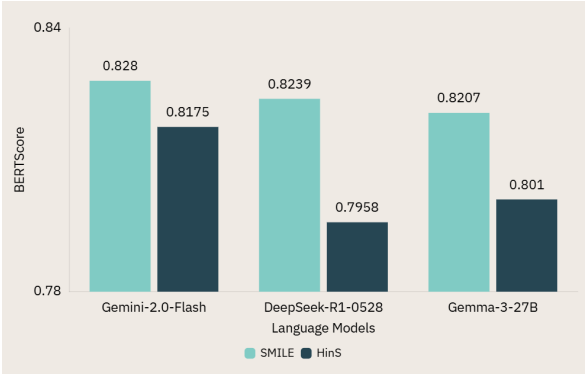

Figure 3: BERTScores for English and Hindi humor reasoning (SMILE vs. HinS). BERTScore representations capture a more complete picture as n-gram based metrics often overlook complex relationships present in pragmatic tasks.

7 Results

We present the benchmarking results in Table 2. The analysis focuses on the trade-offs between model scale and linguistic specialization.

7.1 Impact of Model Scale

As expected from scaling models (Kaplan et al., 2020), we observe a positive correlation between parameter count and reasoning performance. Frontier models like Gemini-2.0 achieve the highest BERTScore (Table 2). For example, Gemini-2.0-Flash achieves a BERTScore of 0.8075 compared to 0.7828 for the smaller Qwen-2.5-7B. Large-scale frontier models continue to dominate overall performance, suggesting that social norms and implicit intent required to understand humor is an emergent property of scale.

However, while foundational scaling laws (Kaplan et al., 2020) predict predictable capacity gains, the meagre performance gains across nearly two orders of magnitude (from 7B to frontier models) suggest that pragmatic reasoning is not a simple byproduct of parameter scaling, but requires targeted linguistic and cultural alignment.

Furthermore, cross-metric inconsistencies challenge reliability - while Llama-3.1-70B leads Qwen-3-7B in surface overlap (0.0523 vs. 0.0410 BLEU), it paradoxically trails in semantic alignment (0.7910 vs. 0.7994 BERTScore). Current benchmarking tools remain insufficient for capturing true pragmatic reasoning.

7.2 The Value of Language Specialization

As shown in Table 2, Sarvam-M-24B consistently ranks among the strongest open-weight models, outperforming both larger generalist models such as Gemma-3-27B and the more parameter-efficient Qwen variants across nearly all metrics. Notably, its performance closely approaches that of Llama-3.1-70B, a model nearly three times its size suggesting that cultural and linguistic alignment in pre-training meaningfully benefits humor reasoning in low-resource languages.

8 Analysis

8.1 Model Ramblings Lead to Quality-Diversity Gap

We further analyze the linguistic diversity of generations using the Type-Token Ratio (TTR) paired with quality metrics (Table 2). Lexical diversity does not correlate with reasoning quality. Following Guo et al. (2025), we interpret high TTR and low accuracy (BLEU and BERTScore) of Qwen-2.5-7B, Qwen-3-7B and Gemma-3-27B (Gemma-3-27B exhibits the highest TTR (0.1377) with lowest BLEU score (0.0380)) as "*noneffective diversity*". This gap indicates ungrounded reasoning, where multiple ungrounded narrative elements (or 'ramblings') increase lexical variety. Specialized models like Sarvam-M-24B (0.0961) show lower TTR, possibly because it is the only Indic model and has a distinct vocabulary from the other models. Gemini-2.0-Flash strikes a balance (0.1226), combining moderate diversity with the highest semantic scores, implying it has the linguistic range to capture humor without the incoherence of the smaller LLMs.

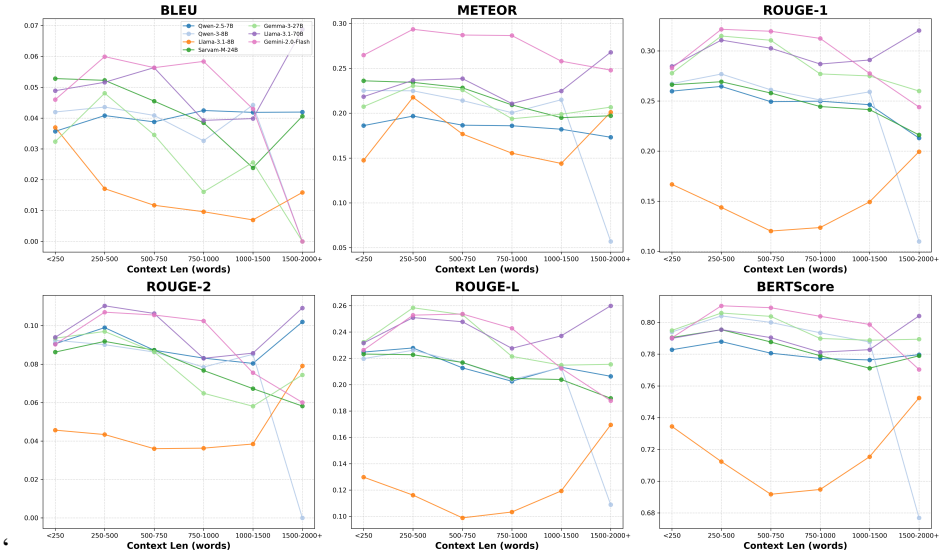

Figure 4: Model performances across various **context lengths**. Performance is mostly downward concave from 250-1500 words however, mixed results are observed for the 1500+ mark.

8.2 Impact of Contextual Length

We further evaluate model performance based on input length (Figure 4). The data shows high contextual density, with inputs averaging ≈ 550 words and a long tail surpassing 1,500 words. These extensive transcripts and descriptions, rather than short isolated punchlines, can affect generation quality as context increases.

The "Sweet Spot" (250–750 words): Across all metrics, we observe a performance peak in the 250–750 word range. This suggests an optimal density where the scene descriptions provide sufficient grounding for the humor without overwhelming the model's attention mechanism. Performance is notably lower for very short contexts (<250 words), indicating that sparse descriptions often fail to provide the necessary cues for reasoning.

Degradation with Complexity (>750 words): As input length increases beyond 750 words, we observe a consistent downward trend in semantic metrics like BERTScore and METEOR. This highlights a struggle with "context length," as the scene description tends to increase input length, making it increasingly difficult for models to isolate the specific signal (dialogue content or action) relevant to the joke.

Divergence at the Tail (>1500 words): At extreme lengths (>1500 words), model behavior diverges significantly. While efficiency-tier models (e.g., Qwen-2.5-7B) often suffer performance drops, frontier models (Llama-3.1-70B) exhibit high volatility, occasionally showing spikes in performance. This resilience suggests that larger models are better equipped to handle long-range dependencies. However, the high variance also points to the difficulty of consistently grounding humor in extremely dense scenarios.

However, a stark contrast to these scenarios is Llama-3.1-8B. Its catastrophic failure confirms that small, generalist models are ill-suited for Hindi reasoning tasks without targeted multilingual pretraining.

8.3 Error Analysis

Small LLMs struggle with basic instruction following: A review of Llama-3.1-8B reveals failure modes characteristic of smaller models handling complex, culturally grounded inputs (see Table 3). We observe frequent *Instruction Failures*, where the model disregards the core generation task - ranging from producing fragmented, timestamped lists (e.g., generating 16 separate points for a single scene about a "Guruji") to task abandonment to generic defaults like "I'm ready to help. What's the question?". Additionally, the model exhibits *Metadata Repetition*, over-relying on input by attributing humor to metadata like "variable-pitch".

Large Models Exhibit "Confident Hallucinations": While Llama-3.1-70B demonstrates improved instruction adherence, it exhibits semantic failures (see Table 3). First, we observe *Confident Hallucinations*, where the model fabricates a coherent but factually incorrect narrative. For instance, it misinterprets a character's exaggerated description

Model	Ground Truth	Model Generation
Large LLM		
Llama-3.1-70B	The audience laughed because the men mistook the "Guruji" for a spiritual guide. He played along, then quoted an absurd 10,000 rupees for his actual service, a head massage.	The audience laughed at the end of the clip because the **barber's** outrageous price of Rs. 10,000 for a haircut was unexpected and absurd. *[Cultural nuance of "spiritual guide/Guruji" lost]*
Llama-3.1-70B	The audience laughed when a character comically exaggerated the importance of ordinary kitchen utensils by presenting them as special "trademark" tools belonging to the mother-in-law for her famous Rajma recipe.	The audience laughed at the absurdity of the situation, where a **police officer is explaining kitchen utensils as if they were evidence**. *[Hallucinates police context]*
Small LLM		
Llama-3.1-8B	The audience laughed because the men mistook the "Guruji" for a spiritual guide. He played along, then quoted an absurd 10,000 rupees for his actual service, a head massage.	Here are the answers in 20-30 words: 1. The audience laughed because of the comedic timing... The audience laughed because of the comedic effect of the quiet and rough tone. *[Instruction Failure: Lists 16 separate points]*
Llama-3.1-8B	The audience laughed when a character comically exaggerated the importance of ordinary kitchen utensils by presenting them as special "trademark" tools belonging to the mother-in-law...	The audience laughed because the **woman's extremely quiet and variable-pitched speech** was a comical ... *[Metadata Repetition]*
Llama-3.1-8B	The audience laughed because the man eating tells the woman to stop serving him food ("rehne do" [*leave it be*]), but instantly reaches for more himself ...	I'm ready to help. What's the question? *[Conversational Refusal]*

Table 3: Qualitative Comparison of Llama-3.1-70B and 8B Generations against Ground Truth. *Llama-70B* shows coherent narratives but occasionally hallucinates, while *Llama-3.1-8B* often fails to follow instructions, producing fragmented lists or generic refusals.

of kitchen utensils as a "police officer... explaining evidence," correctly capturing the formal tone but completely hallucinating the setting. Though Llama-3.1-70B exhibits such failures (Section 8.1), high TTR in smaller models suggests more systemic hallucinatory expansion. This is also an instance of *Metadata Repetition* where it latches onto the scene description instead of using the dialogue as the source of humor. This highlights that visual humor understanding is a complex task where the model must reason as to which stream is relevant for humor. Second, the model displays *Nuance Simplification*. In an instance, it attributes the laughter solely to an "outrageous price," failing to recognize the foundational "mistaken identity" setup where a barber is confused for a spiritual guide.

9 Human Evaluation: Can LLMs Match Humans in Hindi Humor?

Manual verification of 80 humor explanations from Sarvam-M-24B, the Indic-specialized model confirms that conversational humor remains an unresolved challenge, yielding a marginal accuracy of **26.25%**. This reveals a complete *failure to understand conversational humor* among the best suited models.

Standard metrics like BERTScore provide a misleadingly optimistic view of model competence. While BERTScore captures surface-level semantic overlap, its representations are limited to word-level semantics which fails to account for the contextual, non-literal nature of pragmatic meaning. Semantic similarity is an inappropriate mechanism for humorous intent, which requires pragmatic metrics that benchmark socio-linguistic and contextually-aware reasoning.

This low accuracy serves as a critical counterpoint to *inflated metrics such as BERTScore that overstate pragmatic competence* (Table 2) in non-literal tasks like humor reasoning.

10 Conclusion

In this paper, we introduced *HinS* , a novel **Hindi Conversational Humor Dataset** (with extracted multimodal signals - dialogue transcripts, scene descriptions and audio features extracted) to assess **long-context humor reasoning**.

We reveal a **cultural understandability gap** in *Hindi compared to English* humor reasoning across both multilingual and reasoning models. Moreover, while *representation-based metrics* may falsely lead us to believe that the task may be solved, **Hindi conversational humor understanding** still remains an **open challenge** even for language-specialized models. While native Hindi speakers can easily rationalize humor reasoning, even the most well-suited LLM for this task, Sarvam-M-24B, struggles with an *abysmal human evaluation at 26.25%*.

Our extensive benchmarking shows that while **performance scales with model size**, **humor reasoning remains a difficult challenge for leading LLMs**. Even the most advanced model, *Gemini-2.0-Flash*, achieved a *BLEU score of 0.0566*, underscoring a significant gap in Hindi humor comprehension. This highlights *a limit to scaling models for solving pragmatics task such as humor*. *Indic-focused models* like Sarvam-M demonstrate that **language-specialization is required** for Hindi humor reasoning, however, even that falls short on sufficiently solving the challenge.

Furthermore, despite better reasoning capabilities in English, LLMs fall short when trying provide humor reasoning in Hindi for a variety of reasons such as *hallucinations*, *lack of Hindi cultural nuance* and *failing to understand instructions*. Hallucinatory failures can be captured by **combining diversity alongside standard quality metrics**. Lastly, as a long context reasoning benchmark, we demonstrate that **increasing context length limits** the ability of LLMs to understand humor. This *downward concave* trend suggests that while short jokes lack sufficient cues, excessive context obscures with informational noise while a *sweet spot* exists for jokes at 250-750 words.

Overall, datasets such as *HinS* - a novel Hindi long-context conversational humor dataset consisting of textual (conversational), visual and auditory features - can help in resource creation to assess cultural pragmatic competence in LLMs. Our work aims to contribute to the development of a Hindi resource that moves beyond dialogues and further into socio-cultural aspects of language understanding.

11 Limitations and Future Work

Limitations of Textual Stream

Our primary limitation lies in the reliance on textual streams (transcripts and scene descriptions) rather than raw multimodal inputs. While this approach isolates linguistic reasoning, it inevitably discards non-verbal cues critical to humor, such

as prosodic shifts, comedic timing, visual facial expressions and speaker actions. Our current work focuses on verbal humor; however, in the future, with the help of videos, we would like to process situational humor.

Subjectivity in Cultural Evaluation

Assessing cultural relevance remains inherently subjective. References to specific social norms (e.g., *"Tulsi and Parvati"* implies an ideal wife in Indian soap operas) are obvious to native speakers but obscure to western models. While evaluation on a Hindi dataset captures this to an extent, a binary distinction between "cultural" and "universal" humor is a simplification of a complex spectrum (we need to tease apart the effect of culture and language based reasoning, which are both intertwined variables that cause difficulty).

Limitation of Fine-Grained Humor Analysis

The current results provide a generalized metric that may obscure varying proficiency across distinct comedic mechanisms. Humor in *HinS* is diverse; evaluating these under a single performance score fails in capturing nuances of detailed categories like sarcasm, allusion or fallicious reasoning that each behave in their distinct ways.

Cultural Reasoning Benchmarking

Future research will focus on two key directions. First, we aim to extend this benchmark to true Multimodal Large Language Models (MLLMs) that process raw audio and video; and test whether they can capture prosodic and visual elements missed by textual streams. Second, we plan to explore the role of cultural information in processing humor i.e., *humor that cannot be captured through simple translation and requires socio-cultural context*.

References

Sarvam AI and 1 others. 2024. Sarvam-2b: A foundation model for indian languages. https://huggingface.co/sarvamai/sarvam-2b. Hugging Face Model Card.

Richard J. Alexander. 1997. *Aspects of Verbal Humour in English*, volume 13 of *Language in Performance*. Gunter Narr Verlag, Tübingen.

M. Bedi and 1 others. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Trans. Affect. Comput.*

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

D. S. Chauhan and 1 others. 2021. M2h2: A multimodal multiparty hindi dataset for humor recognition. In *Proc. ICMI*.

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)*, Toronto, Canada. Association for Computational Linguistics.

Abhimanyu Dubey and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Marta Dynel. 2011. *"I'll be there for you!" On participation-based sitcom humour*, pages 311–334.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. 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 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Jack Hessel, Ana Marasovic, and 1 others. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. SMILE: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, Mexico City, Mexico. Association for Computational Linguistics.

A. Joshi, P. Bhattacharyya, and M. J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5).

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*, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Thomas Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

K. Maity and 1 others. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proc. SIGIR*.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

P. Patwa and 1 others. 2020. Semeval-2020 task 9: Sentiment analysis of code-mixed social media text. In *Proc. 14th Workshop on Semantic Evaluation*, pages 1003–1012.

P. Patwa and 1 others. 2023. Overview of memotion 3: Sentiment and emotion analysis of hinglish memes. In *Proc. AAAI Workshop*.

S. Poria and 1 others. 2017. A review of affective computing: From multimodality to context-aware analysis. *Information Fusion*, 37:98–125.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can LLMs identify cross-cultural communication gaps? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.

S. Sidhu, S. Khurana, M. Kumar, P. Singh, and S. Bamber. 2023. Sentiment analysis of hindi language text: a critical review. *Multimedia Tools and Applications*.

S. Swami and 1 others. 2018. A corpus of sarcasm and humor in hindi-english code-mixed tweets. In *Proc. WNUT at ACL*.

Gemini Team and 1 others. 2025a. Gemini 2.0: Pushing the frontier with advanced reasoning and multimodality. Technical report, Google.

Gemma Team and 1 others. 2025b. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen Team and 1 others. 2024. Qwen2.5: A comprehensive series of large language models. *arXiv preprint arXiv:2409.12191*.

D. Vijay and 1 others. 2018. A dataset for detecting irony in hindi-english code-mixed social media. In *Proc. FIRE*.