Benchmarking Hindi LLMs: A New Suite of Datasets and a Comparative Analysis

Anusha Kamath, Kanishk Singla, Rakesh Paul, Raviraj Joshi,
Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar
anushak@nvidia.com
NVIDIA

Abstract

Evaluating instruction-tuned Large Language Models in Hindi is challenging due to a lack of high-quality benchmarks, as direct translation of English datasets fails to capture crucial linguistic and cultural nuances.

To address this, we introduce a suite of five Hindi LLM evaluation datasets: **IFEval-Hi, MT-Bench-Hi, GSM8K-Hi, ChatRAG-Hi, and BFCL-Hi.**

These were created using a methodology that combines from-scratch human annotation with a translate-and-verify process. We leverage this suite to conduct an extensive benchmarking of open-source LLMs supporting Hindi, providing a detailed comparative analysis of their current capabilities. Our curation process also serves as a replicable methodology for developing benchmarks in other low-resource languages.

New Datasets

Dataset Name	Count	Method
IFEval-Hi	848	In-house using the foundational framework of English IFEval
MT-Bench-Hi	200	Translated and human evaluated (4 categories); In-house (4 categories)
GSM8K-Hi	1319	Translated and human evaluated (100%)
ChatRAG-Hi	5948	Includes: INSCIT (450), Doc2Dial (498), QuAC, QReCC, TopiocQA, CoQA, HybriDial, SQA, DoQA (Cooking, Travel, Movies), ConvFinQA (500 each). Context: GCP translated, no filtering. Answers and conversation turns: - Used GCP translated data when the back-translated version matched the original (CHRF++ ≥ 90). - Else, used LLM translated data with heuristic filtering to remove poor translations.
BFCL-Hi	2251	Translated (not human evaluated)

Figure 1: Overview of the Hindi evaluation datasets. The test suite consists of Hindi versions of IFEval, MT-Bench, GSM8K, ChatRAG, and BFCL.

IFEval-Hindi

Instruction Following Evaluation (IFEval) benchmark is designed to assess an LLM's ability to adhere to precise instructions.

This dataset includes culture-focused prompts created using content from Indian Wikipedia, featuring 22 main instruction types selected for language relevance and organized into three levels based on difficulty.

Write a short paragraph about X • at least 500 words • no more than 6 sentences • include at least 1 capitalized word	भारत के प्रसिद्ध लेखक और नोबेल पुरस्कार विजेता रवीन्द्रनाथ ठाकुर के बारे में बताइए। • आपका पूरा उत्तर हन्दी में होना चाहिए • कम से कम **पाँच सौ** शब्दों का लेख लिखें। • वाक्यों की संख्या **६** से अधिक न हो। • ~~कहीं भी कम से कम एक शब्द पूरी तरह CAPITAL LETTERS में लिखें।~~

MTBench-Hindi

Multi-Turn Benchmark (MT-Bench), is a standard for evaluating the conversational and reasoning abilities of LLMs.

This dataset was created using a hybrid content creation approach - translating universal technical categories using GCP followed by human verification, while expanding Writing, Roleplay, Humanities, Extraction in-house by human experts.

Question: Pretend yourself to be **Elon Musk** in all the following conversations. Why do we need to go to Mars? Follow up: How do you like dancing? Can you teach me?	Question: मान लीजिए, आप **स्वामी विवेकानंद** हैं, और अब से सभी वार्तालापों में विवेकानंद जी की तरह बोलें। हमें अपने देश के विकास में योगदान क्यों देना चाहिए? Follow up: वैसे, आपको वीडियो गेम खेलना कैसा लगता है?

GSM8K-Hindi

GSM8K (Grade School Math 8K), is a prominent benchmark for assessing the mathematical reasoning of LLMs.

This dataset was developed using a "translate-then-verify" methodology—utilizing GCP for initial machine translation of the English mathematical problems followed by rigorous human expert review.

Tim decides to light off some fireworks for the fourth of July. He buys a package of fireworks worth $400 and **another pack worth twice that much.** He gets a 20% discount on them. He also buys a finale firework that costs $150. How much did he spend in total?	Before Correction: टिम ने चौथी जुलाई के लिए कुछ आतिशबाजी जलाने का फैसला किया। **वह $400 के पैकेज और उतनी ही कीमत के दो पैक खरीदता है।**... After Correction: टिम ने चौथी जुलाई के लिए कुछ आतिशबाजी जलाने का फैसला किया। **वह 400 डॉलर मूल्य का एक पैकेट पटाखों का खरीदता है तथा उससे दुगुना मूल्य का एक और पैकेट खरीदता है।** ....

ChatRAG-Hindi

ChatRAG Bench, is a benchmark for evaluating conversational question-answering using documents and retrieved context. It original incorporates ten diverse datasets, including Doc2Dial, QuAC, and ConvFinQA.

This dataset was created by translation of context passages using GCP and a carefully designed filtration and processing layer for questions and answers to filter the high quality samples after translation.

Question
क्या आपको पता है कि इरकुत्स्क में वीज़ा प्राप्त करना आसान है या अलमाटी में?

Answer
अलमाटी में एक मंगोलियाई दूतावास है, लेकिन यह शहर से काफी दूर है, और वहाँ पहुँचना परेशानी भरा है।

Context
अंत में, मैंने इरकुत्स्क को चुना। मेरा तर्क: अलमाटी में एक मंगोलियाई दूतावास है, लेकिन यह शहर से काफी दूर है, और वहाँ पहुँचना परेशानी भरा है। इसके अलावा, कज़ाख पुलिस मुझे थोड़ी चिंतित करती है, और जब वे आसपास होते हैं तो पासपोर्ट के.....

BFCL-Hindi

Berkeley Function-Calling Leaderboard (BFCL v2), benchmark designed to evaluate the ability of LLMs to call functions or tools.

This dataset is created to evaluate the model's ability to understand a Hindi query and map it to a pre-defined English language tool. The conversational history is translated into Hindi using the GCP translation service and functions are used as it is.

Question:
"Book a one-way economy flight from San Francisco to New York next Friday after 5 PM for two adults, window seats if available."

Functions:
search_flights*(origin, destination, date, ...)*
book_flight*(flight_id, passengers, ...)*

Translated Question:
अगले शुक्रवार शाम 5 बजे के बाद सैन फ्रांसिस्को से न्यूयॉर्क के लिए दो वयस्कों हेतु एक वन-वे (एक तरफा) इकोनॉमी फ्लाइट बुक करें, और यदि उपलब्ध हों तो विंडो सीट लें।

Results and Resources

Model	Size	MT-Bench-Hi	BFCL-Hi	GSM8K-Hi	IFEval-Hi	ChatRAG-Hi
SLMs						
Gemma-2-2b-it	2B	4.37	32.96	26.99	38.92	29.89
Llama-3.2-3B-Instruct	3B	5.14	33.81	40.11	40.80	32.60
Nemotron-Mini-4B-Instruct	4B	3.44	-	32.22	36.08	27.32
Nemotron-4-Mini-Hindi-4B-Instruct	4B	6.01	**52.82**	47.31	51.65	36.07
Llama-3.1-8B-Instruct	8B	6.44	31.23	61.33	48.82	38.03
Aya-expanse-8b	8B	6.58	36.56	**64.52**	42.92	30.15
Gemma-2-9b-it	9B	**7.37**	50.51	64.44	**61.79**	**40.97**
Krutrim-2-instruct	12B	6.31	26.88	56.56	59.32	37.48
LLMs						
GPT-OSS-20B (reasoning low)	21B	8.51	54.60	80.64	69.04	26.16
Mistral-Small-3.2-24B-Instruct-2506	24B	7.83	41.45	77.55	66.89	37.92
Sarvam-M (reasoning off)	24B	8.25	48.60	82.30	71.64	40.14
Gemma-3-27b-it	27B	8.31	**62.42**	78.12	67.72	45.23
GPT-OSS-120B (reasoning low)	117B	**8.70**	61.26	**93.41**	**73.86**	29.85
Qwen3-235B-A22B-FP8 (reasoning off)	235B	8.10	59.88	89.69	68.11	32.47
Llama-3.1-405B	405B	7.17	49.53	86.27	68.66	**47.46**
LLMs (Reasoning)						
GPT-OSS-20B (reasoning medium)	21B	8.43	63.26	83.41	72.01	29.16
GPT-OSS-20B (reasoning high)	21B	8.23	64.77	83.44	72.11	32.39
Sarvam-M (reasoning on)	24B	8.60	59.53	84.40	74.06	**37.13**
GPT-OSS-120B (reasoning medium)	117B	**8.79**	**66.19**	95.93	76.69	30.80
GPT-OSS-120B (reasoning high)	117B	8.70	64.90	**96.27**	**76.80**	31.82

SLMs (<20B params): Gemma-2-9B-it: Best overall on MT-Bench-Hi, IFEval-Hi, ChatRAG-Hi.

LLMs (>20B params): GPT-OSS-120B: Best on MT-Bench-Hi, GSM8K-Hi, IFEval-Hi.

This work fills the evaluation gap for Hindi instruction-tuned LLMs by introducing five culturally and linguistically robust benchmarks curated through a hybrid human-plus-translation framework adaptable to other languages.

Resources

- https://huggingface.co/datasets/nvidia/IFEval-Hi
- https://huggingface.co/datasets/nvidia/MT-Bench-Hi
- https://huggingface.co/datasets/nvidia/GSM8K-Hi
- https://huggingface.co/datasets/nvidia/ChatRAG-Hi
- https://huggingface.co/datasets/nvidia/BFCL-Hi
- **Full paper:** https://arxiv.org/pdf/2508.19831