# 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**
NVIDIA
{anushak, kanishks, rapaul, ravirajj, uvaidya,
schauhan, nwartikar}@nvidia.com

## Abstract

Evaluating instruction-tuned Large Language Models (LLMs) 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.

## 1 Introduction

The rapid expansion of Large Language Models (LLMs) necessitates the development of robust and reliable evaluation methodologies (Liang et al., 2022; Srivastava et al., 2023). As these models are integrated into a wide range of applications, a rigorous assessment of their capabilities, limitations, and safety is paramount (Achiam et al., 2023; Wang et al., 2023). Although the initial focus of evaluation has been predominantly on English, a model's global utility is contingent upon its performance across diverse linguistic and cultural contexts (Singh et al., 2024c). The evaluation of non-English LLMs is therefore essential, not only for ensuring equitable technological access but also for understanding the extent to which these models capture the distinct complexities inherent in different languages, an undertaking that goes beyond mere translation (Bender et al., 2021).

The evaluation landscape for English LLMs is well-established, featuring a comprehensive suite of benchmarks targeting a spectrum of model capabilities. For foundational "base" models, benchmarks assess commonsense reasoning, such as HellaSwag (Zellers et al., 2019) and Winogrande (Sakaguchi et al., 2021), factual accuracy with TruthfulQA (Lin et al., 2022), and broad multi-task knowledge with MMLU (Hendrycks et al., 2020; Wang et al., 2024; Singh et al., 2024b). Specialized datasets evaluate capabilities like mathematical reasoning on GSM8K(Cobbe et al., 2021) and code generation with HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). Furthermore, the advent of interactive, instruction-following models has spurred the creation of benchmarks to assess conversational quality on MT-Bench (Zheng et al., 2023), fidelity to complex instructions with IFEval (Zhou et al., 2023), and the ability to execute tool or function calls correctly on BFCL (Patil et al.). These datasets have collectively become the standard for evaluating the performance of state-of-the-art models in English.

In recent years, significant progress has been made in developing evaluation resources for Indic languages, typified by benchmarks such as IndicGLUE (Kakwani et al., 2020), MILU (Verma et al., 2025), IndicMMLU-Pro (Sankalp et al.), and IndicGenBench (Singh et al., 2024a). These resources have been instrumental in assessing the core capabilities of pre-trained base models across numerous languages of the Indian subcontinent (Joshi et al., 2024). Despite this progress, the existing benchmarks primarily target pre-trained base models, leaving a noticeable gap in resources for assessing the capabilities of instruction-tuned models. Consequently, benchmarks for critical skills like instruction following, conversational ability, and function calling, such as Hindi versions of IFEval, MT-Bench, and BFCL, are largely unavailable publicly.

A common methodology to address this gap involves the direct translation of existing English benchmarks. This approach, however, presents considerable challenges, as automated translation



*Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)*, pages 52–68

| Dataset Name | Count | Method |
|---|---|---|
| IFEval-Hi | 848 | In-house |
| 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 | Translated, filtered, and human-evaluated (5%). Includes: INSCIT (450), Doc2Dial (498), QuAC, QReCC, TopiocQA, CoQA, HybriDial, SQA, DoQA (Cooking, Travel, Movies), ConvFinQA (500 each).<br>Context: GCP translated, no filtering.<br>Answers and conversation turns:<br>- Used GCP translated data when the back-translated version matched the original (CHRF++ $\geq$ 90).<br>- Else, used LLM translated data with heuristic filtering to remove poor translations. |
| BFCL-Hi | 2251 | Translated (not human evaluated) |

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

frequently fails to preserve the linguistic subtleties and cultural context integral to the target language. This process can yield datasets that are linguistically incongruous or culturally irrelevant, thereby diminishing the validity and reliability of the evaluation. Such benchmarks often test a model's ability to comprehend translated English rather than its native fluency and instruction fidelity.

To address these deficiencies, this paper introduces Hindi versions of five widely-used and comprehensive benchmarks: IFEval-Hi, MT-Bench-Hi, GSM8K-Hi, ChatRAG-Hi, and BFCL-Hi. We developed these datasets using a process that combined direct human creation with a translate-and-verify workflow, ensuring high linguistic and cultural relevance. A summary of the final dataset sizes and curation methods is presented in Table 1. Furthermore, we utilize this new suite to conduct a comprehensive benchmarking of several prominent, publicly available LLMs based on foundational models, including Llama, Gemma, and Nemotron. This work contributes a valuable, high-quality evaluation suite for Hindi to the research community and presents a comparative analysis that offers critical insights into the current capabilities of Hindi language models.

The main contributions of our work are as follows:

- We introduce a suite of five new, high-quality benchmarks (IFEval-Hi[1], MT-Bench-Hi[2], GSM8K-Hi[3], ChatRAG-Hi[4], and BFCL-Hi[5]) for evaluating instruction-tuned LLMs in Hindi and detail the curation process developed for their creation.
- We present a comprehensive benchmark of prominent, publicly available LLMs on this new suite, providing the first robust comparative analysis of their capabilities in Hindi. Our findings show that while specialized models exhibit strength in specific tasks, Gemma-2-9b-it in the SLM class and GPT-OSS-120B in the LLM class emerge as the most capable general-purpose models.

[1] https://huggingface.co/datasets/nvidia/IFEval-Hi
[2] https://huggingface.co/datasets/nvidia/MT-Bench-Hi
[3] https://huggingface.co/datasets/nvidia/GSM8K-Hi
[4] https://huggingface.co/datasets/nvidia/ChatRAG-Hi
[5] https://huggingface.co/datasets/nvidia/BFCL-Hi

## 2 Related Work

Recent years have witnessed notable progress in the evaluation of multilingual and low-resource language models, with a particular focus on Indic languages. Foundational efforts, such as IndicGLUE (Kakwani et al., 2020) and IndicXTREME (Doddapaneni et al., 2023), established the initial groundwork by adapting the GLUE paradigm for major Indic languages. These benchmarks provided a broad suite of Natural Language Understanding (NLU) tasks, including classification, entailment, and named entity recognition, which proved instrumental in assessing the foundational capabilities of models across multiple Indic languages, including Hindi.

Building upon these foundations, subsequent benchmarks like MILU (Verma et al., 2025) introduced more challenging and culturally grounded tasks. MILU, is a large-scale benchmark comprising approximately 80,000 multiple-choice questions derived from Indian competitive examinations. By emphasizing India-specific domains such as local governance, arts, and history, MILU underscores the importance of cultural context in evaluation, an element often diluted in directly translated datasets. Specialized datasets like IndicQuest (Rohera et al., 2024) have been developed to evaluate the factual knowledge of Indic LLMs. In parallel,

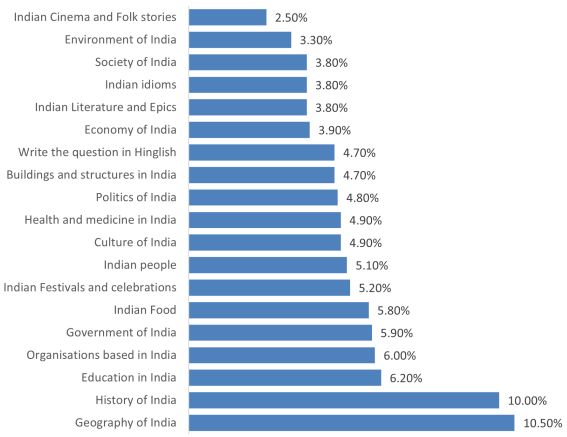


Figure 1: Distribution of samples by Indian cultural themes in the IFEval-Hi dataset.

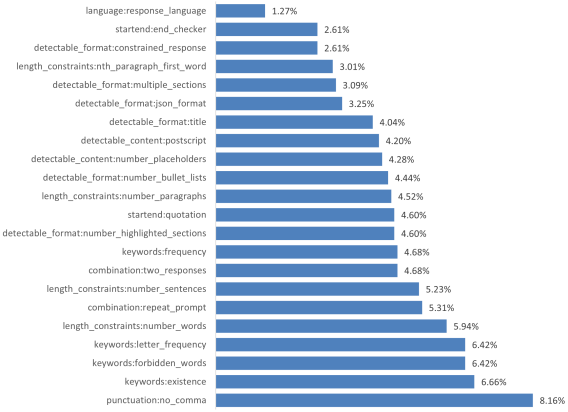


Figure 2: Distribution of verifiable instruction types within the IFEval-Hi dataset.

benchmarks such as IndicSQuAD (Endait et al., 2025) and IndicQA (Singh et al., 2025) have addressed extractive and abstractive question answering.

More recently, the field has shifted toward multi-task and generative evaluation. Benchmarks like the IndicGenBench suite (Singh et al., 2024a) and IndicMMLU-Pro (Sankalp et al.) now assess complex reasoning, creative understanding, and instruction-following, demonstrating a move beyond traditional NLU paradigms. This trend is further reflected in the Okapi (Lai et al., 2023), which translated key English benchmarks into numerous languages, and the development of Global MMLU (Singh et al., 2024b), which extends evaluation to more diverse cultural contexts.

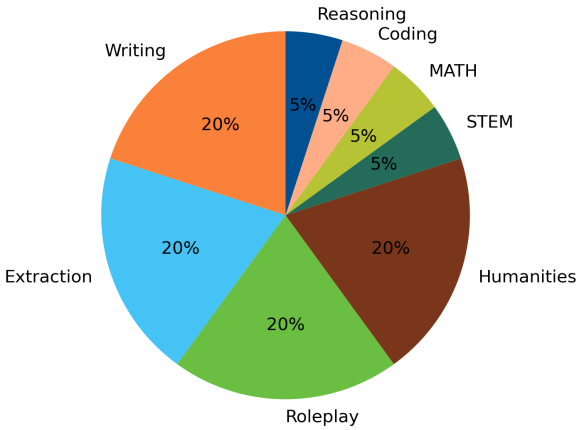


Figure 3: Category distribution in MT-Bench-Hi, adapted with Indian cultural themes to increase focus on culturally relevant instructions.

## 3 Dataset Curation

This section describes the process used to curate the Hindi versions of popular English benchmark datasets. Sample examples from each dataset are shown in Figure 4.

### 3.1 IFEval-Hi

The creation of IFEval-Hi is based on the English Instruction Following Evaluation (IFEval) benchmark, which is designed to rigorously assess an LLM's ability to adhere to precise instructions. The original English IFEval is structured around 25 distinct and objectively verifiable instruction types, such as "insert a word at a specific position" or "reverse the first paragraph". This structure ensures a reliable and scalable evaluation process. To test models with increasing difficulty, the prompts are organized into three complexity levels: Single, Double, and Triple Instructions, requiring the model to execute one, two, or three distinct commands within the same prompt, respectively.

Our curation process for IFEval-Hi was a systematic adaptation of this framework to an Indian cultural and linguistic context. The core of this process involved retaining the 22 verifiable instruction types as a structural framework while replacing the generic content of the English prompts with themes relevant to India. Some categories that are not relevant to Hindi, such as "Change Cases," were dropped. The thematic content was sourced from comprehensive categories on Wikipedia related to India, covering a wide range of topics including Indian history, philosophy, festivals, art forms, and social norms. The distribution of cultural themes

| | |
|---|---|
| **IFEval-Hi** | भारत की शिक्षा प्रणाली के सुधार के बारे में अपनी राय रखें। आपका पूरा उत्तर हिन्दी में होना चाहिए और उसमें "श" अक्षर कम से कम ६ बार आना चाहिए। |
| **MT-Bench-Hi** | खुद को रत्न टाटा के रूप में प्रस्तुत करें और अगले संवाद में उनके जैसा बोलने की कोशिश करें। भारत में सौर ऊर्जा में निवेश क्यों करना चाहिए, और क्या यह देश के भविष्य के लिए महत्वपूर्ण है?<br>*Follow up:* भारत में सौर ऊर्जा के क्षेत्र में निजी क्षेत्र और सरकार को मिलकर किस प्रकार की नीतियाँ अपनानी चाहिए, ताकि यह क्षेत्र और भी तेजी से विकसित हो सके? |
| **GSM8K-Hi** | एक टोकरी में 25 संतरे हैं, जिनमें से 1 खराब है, 20% कच्चे हैं, 2 खट्टे हैं और बाकी अच्छे हैं। कितने संतरे अच्छे हैं? |
| **ChatRAG-Hi** | *Question:* जापानी ट्रेन शिष्टाचार: क्या शिंकान्सेन ग्रीन कार में शिशु ले जाना स्वीकार्य है?<br>*Answer:* यह स्वीकार्य है। हालांकि, यह शिष्टाचार है कि यदि बच्चा शोर मचाना शुरू कर दे तो आप उसे दरवाजे के बाहर "डेक" क्षेत्र में ले जाएं।<br>*Context:* हां, यह स्वीकार्य है। हालांकि, यह शिष्टाचार है कि यदि बच्चा शोर मचाना शुरू कर दे तो आप उसे दरवाजे के बाहर "डेक" क्षेत्र में ले जाएं (जहां बाथरूम, टेलीफोन, वेंडिंग मशीन आदि हैं)। ... |
| **BFCL-Hi** | Prompt: इसे 20 डिग्री घुमाएं और क्षैतिज रूप से पलटें<br>*Function Names:* flipImageAction, rotateImageAction, removeBackgroundAction, getRecommendationsAction, resizeImageAction |

Figure 4: Representative examples from five Hindi evaluation datasets curated in this study.

is detailed in Figure 1, while the breakdown across verifiable instruction categories is presented in Figure 2.

New prompts were carefully created by a team of five annotators over a ten-week period. To ensure that the newly created Indian-themed prompts were both culturally relevant and objectively verifiable, annotators were provided with examples for each instruction type. For instance, when the instruction theme is "Geography of India" and the instruction category is a letter frequency constraint, such as requiring a certain Hindi letter to appear at least three times, the annotator crafts an instruction that incorporates both the theme and the explicit constraint. Specific sample is shown in Figure 4. To ensure that IFEval-Hi could be used as a direct benchmark against its English counterpart, the evaluation metrics and constraints for each of the 22 instruction categories were directly mirrored, along with the three levels of complexity. This significant human-in-the-loop effort resulted in a final dataset comprising 848 high-quality, culturally resonant samples. The annotation process is described in further detail in Appendix A.2.

### 3.2 MT-Bench-Hi

MT-Bench-Hi is the Hindi adaptation of the English Multi-Turn Benchmark (MT-Bench), a standard for evaluating the conversational and reasoning abilities of LLMs in extended dialogues. The original benchmark consists of 80 high-quality, multi-turn questions designed to test key capabilities such as maintaining context, response accuracy, and instruction following. It employs an "LLM-as-a-Judge" approach, where a powerful model like GPT-4o scores all responses on a 1-10 scale using two distinct methods: for reference-free categories (STEM, Writing, Roleplay, Humanities, Extraction), responses are scored directly, while for categories with reference answers (Reasoning, Math, Coding), they are evaluated via pairwise comparison against the reference answer.

The curation of MT-Bench-Hi was a detailed adaptation process designed to make the benchmark culturally and contextually relevant for India. We adopted a hybrid approach to content creation. For universal technical categories (STEM, math, reasoning, coding), questions were translated from English to Hindi using GCP and subsequently underwent thorough human evaluation to verify accuracy and intent. For categories requiring deep cultural contextualization (Writing, Roleplay, Humanities, Extraction), questions were created from scratch by human specialists to ensure the prompts were authentically Indian. Figure 3 illustrates the final distribution of samples across these categories.

To maintain high standards, annotators were provided with reference examples from the English MT-Bench and guided through a specialized interface. A key quality assurance step involved showing annotators sample responses from a high-performing model (e.g., GPT-4o) to help them craft prompts that could effectively test advanced capabilities in an Indian context. The evaluation framework was aligned with the original's "LLM-as-a-Judge" methodology. To ensure consistency, we maintain the same format as the original dataset; for categories that include a reference answer, we

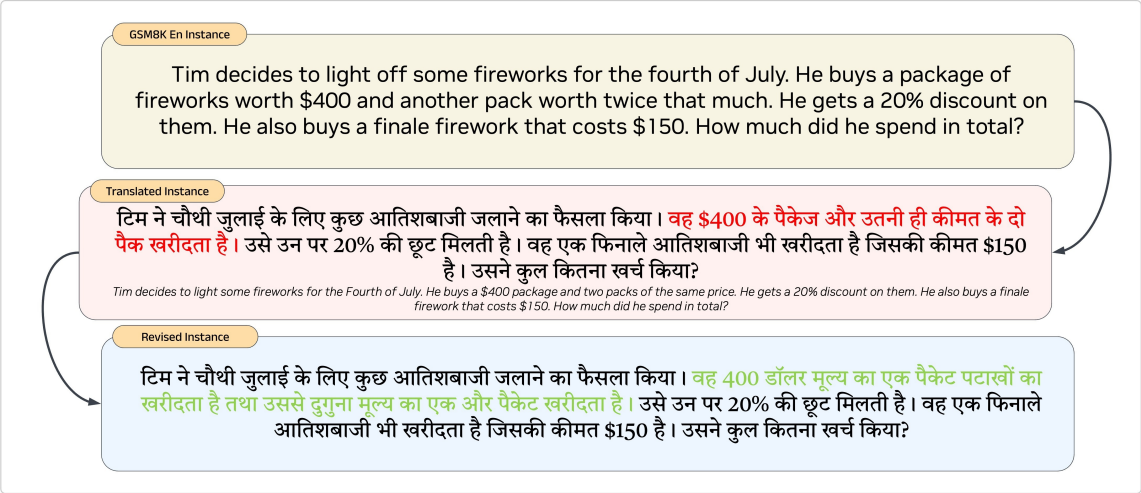


Figure 5: A sample GSM8K question highlighting a translation mistake in Hindi (red), the corrected version (green), and the corresponding English line (yellow), showcasing the process of identifying and fixing language conversion errors manually.

retain the original English reference answer during evaluation. Aligning with the original MT-Bench setup, we employ direct, single-answer evaluation for reference-free, subjective categories (e.g., Writing, Roleplay) and pairwise comparison against a reference answer for categories with objective solutions (e.g., Math, Coding, Reasoning). The annotation process is described in further detail in Appendix A.3.

### 3.3 GSM8K-Hi

The foundation for this dataset is the English GSM8K (Grade School Math 8K), a prominent benchmark for assessing the mathematical reasoning of LLMs. Directly translating the dataset into Hindi risks altering the underlying mathematical logic, particularly in problems with comparative constructs such as 'twice that amount' or '10 less than half the age of'. However, crafting linguistically diverse math problems that require multi-step solutions demands significant expertise in both mathematics and language structure. This process involves considerable time and effort from human annotators with domain expertise, making it a resource-intensive endeavor. Therefore, to balance quality with feasibility, we opted for a two-step "translate-then-verify" methodology.

The process began with machine translation of the original English problems (including GSM8K system prompt) into Hindi using GCP. Human annotators were then provided with both the Hindi translation and the original English text for reference. Their primary task was to evaluate the Hindi translation for correctness and suggest modifications to ensure linguistic accuracy and contextual appropriateness. This verification stage proved to be essential, as annotators flagged approximately 10% of the machine-translated data for inaccuracies. These instances were subsequently reviewed and corrected through close collaboration between our development team and the annotators, ensuring the final dataset maintained high quality. To illustrate this process, Figure 5 shows a sample error alongside its correction. To ensure consistent benchmarking, GSM8K-Hi is evaluated using the LM-Eval-Harness, the same framework employed for the original English dataset. The annotation process is described in further detail in Appendix A.4.

### 3.4 ChatRAG-Hi

ChatRAG-Hi is the Hindi version of ChatRAG Bench, a benchmark for evaluating conversational question-answering using documents and retrieved context. The original incorporates ten diverse datasets, including Doc2Dial, QuAC, and ConvFinQA. Adapting this composite dataset posed challenges due to the varied structures of its subsets, which range from extensive contexts to single-word answers.

Our curation process involved a differential translation strategy. The extensive context passages were translated using GCP without subsequent filtering. For the more sensitive answers and conversation turns, we adopted a two-tiered approach. We first used GCP and validated the output by back-

| 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 |

Table 2: Performance of various LLMs on Hindi benchmarks. MT-Bench-Hi is scored on a scale of 1-10 using an LLM-as-a-judge approach. BFCL-Hi, GSM8K-Hi, and IFEval-Hi report accuracy on a 1-100 scale, while ChatRAG-Hi reports the F1-Score. The highest score in each column is highlighted in bold.

translating it to English. If the back-translated text matched the original with a high degree of fidelity (CHRF++ score >= 90), the GCP translation was retained. In cases where the CHRF++ score was low, which often occurred with very short text segments (1–3 words) where GCP lacks sufficient contextual cues, the GCP translation was discarded. To overcome this, we employed an LLM for these segments, providing it with the broader GCP-translated Hindi context alongside the original short English answer to generate a more accurate and contextually appropriate Hindi equivalent. This LLM-generated (Llama-3.1-405B) data was then subjected to heuristic filtering to remove poor-quality outputs. This hybrid methodology was designed to maximize accuracy across different text types. To ensure overall quality, approximately 10% of the final Hindi data underwent human verification, which confirmed the high fidelity of the translations, with the error rate across subsets remaining within 1-5%.

### 3.5 BFCL-Hi

BFCL-Hi is the Hindi adaptation of the Berkeley Function-Calling Leaderboard (BFCL V2), a benchmark designed to evaluate the ability of LLMs to call functions or tools. The original dataset comprises diverse function-calling scenarios, including simple, multiple, and parallel calls. It also includes relevance and irrelevance categories to assess a model's ability to determine if the provided tools are appropriate for a given query.

The dataset is structured in a JSON format where each entry contains a conversation history and an array of available functions, defined with names, descriptions, and parameter schemas. To create BFCL-Hi, we translated the conversational history into Hindi using the GCP translation service. Crucially, the function calls themselves, including their names, descriptions, and parameter details, were retained in their original English format. This hybrid approach tests the model's ability to understand a Hindi query and map it to a predefined English-language tool. However, to make the dataset more relevant for fully localized use cases, the function parameters should also be translated into Hindi, which we leave as a task for future work. The ground truth for simple, multiple, and parallel categories remained unchanged from the English version. The relevance and irrelevance categories do not include ground truth, as they are designed to verify whether the model correctly attempts a function call. Evaluation is performed using the BFCL Abstract Syntax Tree (AST) methodology to ensure a thorough and accurate analysis.

## 4 Results and Discussion

This section presents and analyzes the performance of a diverse set of publicly available, instruction-tuned Small Language Models (SLMs) and Large Language Models (LLMs) on our newly developed Hindi benchmark suite, with detailed results presented in Table 2. The models evaluated include representatives from prominent families such as Google's Gemma, Meta's Llama, OpenAI's GPT-OSS, NVIDIA's Nemotron, Qwen, and Sarvam, alongside other notable multilingual models.

Among the SLMs, the results reveal a competitive landscape. Gemma-2-9b-it provides the best all-around performance, securing the highest scores on MT-Bench-Hi, IFEval-Hi, and ChatRAG-Hi. Aya-expanse-8b secures the best score on GSM8K-Hi. The value of targeted, language-specific training is highlighted by Nemotron-4-Mini-Hindi-4B-Instruct, which leads significantly on BFCL-Hi.

In the LLM category (models with > 20B parameters), GPT-OSS-120B demonstrates standout performance by achieving the best scores on MT-Bench-Hi, GSM8K-Hi, and IFEval-Hi. Other models show specialized strengths: Gemma-3-27b-it achieves the highest score on BFCL-Hi, while the largest model, Llama-3.1-405B, excels on ChatRAG-Hi. However, it is worth noting that GPT-OSS may have an inherent advantage due to its reasoning mode, even though we set it to a low level for a fairer comparison, and the potential for the GPT-4o judge to be biased towards a sibling OpenAI model also warrants further investigation.

Furthermore, activating the dedicated reasoning modes in models like GPT-OSS and Sarvam-M provides a substantial performance uplift on complex tasks like BFCL, GSM8K, and IFEval. With these capabilities enabled, GPT-OSS-120B achieves the top scores across multiple benchmarks, highlighting the value of reasoning models for Hindi.

In summary, while specialized models show strength in specific tasks, Gemma-2-9b-it in the SLM class and GPT-OSS-120B in the LLM class emerge as the most capable general-purpose models. The distribution of top scores across different models highlights that no single model is best for all tasks. This analysis also indicates that model size is not the sole determinant of performance, a point reinforced by both the 8B Aya model outperforming larger SLMs on GSM8K-Hi and the competitive results of Sarvam-M, which was post-trained on Indic languages. These findings suggest that architectural choices and targeted training data are crucial for developing specialized capabilities for the Hindi language.

## 5 Conclusion

In this work, we addressed the critical gap in evaluation resources for instruction-tuned Hindi LLMs by introducing a new suite of five culturally and linguistically robust benchmarks. Our hybrid curation methodology, combining careful human-centric creation with a translate-and-verify process, provides a valuable framework for developing similar resources in other languages. Our evaluation of various public LLMs supporting the Hindi language revealed a competitive landscape where different models exhibit specialized strengths in reasoning, conversation, and function calling. This suite enables a more nuanced assessment of Hindi LLMs, supporting the broader goal of fostering more equitable and capable multilingual AI systems.

## Limitations

We acknowledge certain limitations in our work. While our benchmark suite is comprehensive, it does not encompass every possible instruction type or conversational scenario. The use of an "LLM-as-a-Judge" for MT-Bench-Hi carries inherent biases, particularly as the judge model's proficiency in evaluating nuanced Hindi content is not guaranteed. Furthermore, datasets developed through translation, despite human verification, could be improved with full human curation to better capture cultural and linguistic subtleties. Future work could expand the scope of these benchmarks and explore alternative evaluation methodologies.

## Acknowledgements

This work would not have been possible without contributions from many people at NVIDIA. To mention a few: Priyanka Chavan, Shweta Dash, Asmita Kadwe, Nilesh Chauhan, Saloni Pipada, Shrutika Marke, Priyanka Patil, Ninad Nemade, Muskan Grewal, Aditya Patil, and Noopur Mishra.

The authors also acknowledge the use of AI-powered language models to assist with the editing and revision of this manuscript.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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.

Sharvi Endait, Ruturaj Ghatage, Aditya Kulkarni, Rajlaxmi Patil, and Raviraj Joshi. 2025. Indicsquad: A comprehensive multilingual question answering dataset for indic languages. *arXiv preprint arXiv:2505.03688*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, 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.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 982–988.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

KJ Sankalp, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio,

Wei Qi Leong, Yosephine Susanto, and 1 others. 2024b. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, and 1 others. 2024c. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. 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 10076–10132.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

# A Appendix

## A.1 Annotator Team and Process

The human-centric curation and verification tasks were conducted by a team of five specialists. These individuals were employees of our organization, compensated fairly for their work, and were selected for their proficiency in Hindi as either a first or second language. Representing various regions across India, they possessed strong reading and writing skills and a solid understanding of cultural nuances.

The primary tool used for annotation was SuperAnnotate, which provided an intuitive interface for our workflow. This platform allowed for the efficient sharing of examples, processing of results using Python scripts, and performance of quality assurance (QA). The data underwent periodic QA and development checks to ensure alignment with project requirements. To maintain high levels of creativity and productivity, the specialists worked in focused sessions of 2–3 hours per day.

## A.2 IFEval-Hi Curation Process

The annotation procedure for IFEval-Hi was highly structured and communicated to annotators through a comprehensive guidelines document. The process was organized into sequential stages of increasing complexity, beginning with cases that contained a single verifiable instruction, followed by stages with two and then three instructions. Each test case was assigned a predefined Indian theme and instruction category to ensure a balanced distribution of scenarios.

The annotation workflow for each case involved several key components:

- Reference Sample: Annotators were provided with a developer-generated sample in Hindi that incorporated the verifiable instruction, serving as a clear reference for the task requirements.
- Annotation Interface: A dedicated text box was provided for annotators to formulate their questions based on the assigned theme and instruction category, with a separate comments box for any necessary clarifications with the quality control developer.
- Evaluation Parameters: The parameters required for automatic evaluation were also systematically recorded, aligning with the stan-

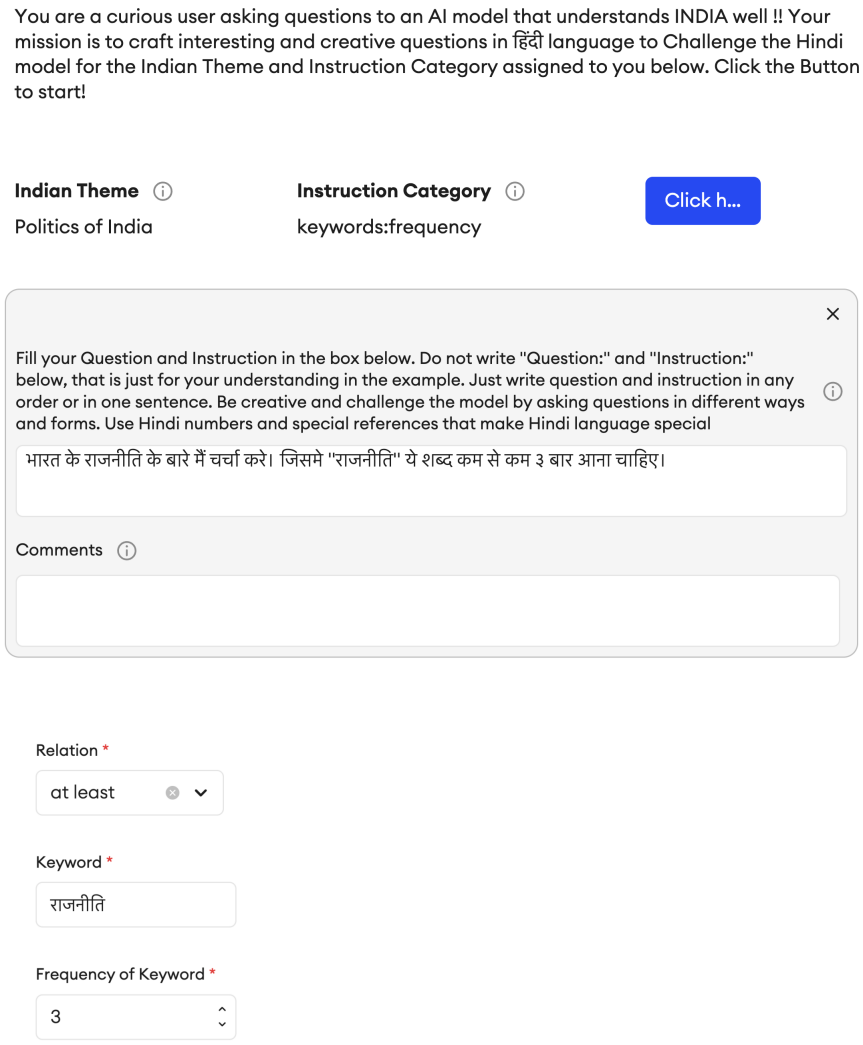


Figure 6: Illustration of the annotation interface used to curate the IFEval dataset, displaying the guidelines and example instructions provided to annotators.

dards of the English dataset. Annotators received detailed guidelines for these parameters, with the Hindi reference sample serving as a practical model.

- Review and Feedback Loop: A weekly review of approximately 50% of submitted samples was conducted by developers. Any cases requiring revision were returned to annotators with specific feedback in the comments section, ensuring a consistent feedback loop and high-quality output.

Sample annotation UI screens are shown in Figures 6 and 7. Some examples from the dataset are shown in Figure 8.

### A.3 MT-Bench-Hi Curation Process

The curation of the MT-Bench-Hi dataset, while presenting distinct challenges, benefited from the procedural learnings established during the IFEval-Hi creation. Specialists were guided by supplementary instructions tailored to the specific demands of creating multi-turn conversational benchmarks. This process was designed to help annotators understand the evaluation procedure and produce high-quality, contextually relevant samples.

The workflow for each test case provided annotators with a comprehensive view of the task:

- Original English Sample: Annotators were given an original question and follow-up from the MT-Bench dataset as a reference.

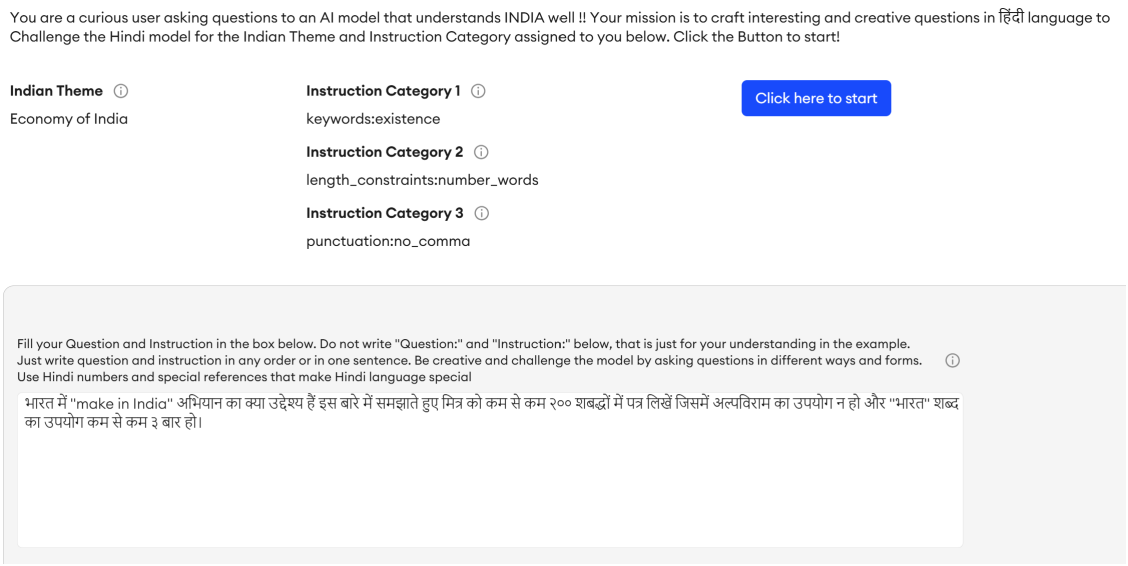


Figure 7: Example entry from the Hindi IFEval, curated at three levels of complexity: Single Instruction, Double Instruction, and Triple Instruction.

| **IFEval-Hi** |
|---|
| भारत की शिक्षा प्रणाली के सुधार के बारे में अपनी राय रखें। आपका पूरा उत्तर हिन्दी में होना चाहिए और उसमें "श" अक्षर कम से कम ६ बार आना चाहिए। |
| भारत के प्रसिद्ध नृत्यों पर दो लेख लिखें और दोनों लेखों को ****** से अलग करें। |
| भारत के प्रमुख वन्यजीव अभयारण्यों, राष्ट्रीय उद्यानों तथा संरक्षण में उनकी भूमिका के बारे में बताइए। ध्यान रखें कि आप इसमें अल्पविराम का प्रयोग न करें और प्रजाति एवं संरक्षण जैसे शब्दों को शामिल करना न भूलें। |
| भारत में अंग्रेज़ों ने कितने वर्षों तक राज किया, इस पर एक बड़ा लेख लिखें। इस लेख का शीर्षक दोहरे कोणीय कोष्ठक में होना चाहिए, जैसे <<ब्रिटिश राज>>। |
| केंद्रीय प्रदूषण नियंत्रण बोर्ड का गठन कब हुआ था और इसके कार्य क्या हैं, इस विषय पर २० से ज़्यादा वाक्यों और न्यूनतम २०० शब्दों में एक लेख लिखें। आपके लेख का शीर्षक दोहरे कोणीय कोष्ठक में होना चाहिए, जैसे <<प्रदूषण नियंत्रण>>। |

Figure 8: Examples from the IFEval-Hi benchmark.

- Model Response Example: The corresponding model-generated response for the English sample was included.
- Evaluation Insight: The AI judge's rating and judgment for that response were also provided, offering annotators direct insight into the evaluation criteria and performance expectations.

Using this framework, annotators reviewed the initial English question and its follow-up, then crafted analogous questions contextualized for Indian settings. To ensure quality and adherence to guidelines, 50% of the newly created samples were subject to a weekly review by a developer. This structured approach equipped annotators to produce high-quality, contextually appropriate samples for the MT-Bench-Hi dataset. The sample annotation UI screen is shown in Figure 9. Some examples from the dataset are shown in Figure 10.

### A.4 GSM8K-Hi Curation Process

By the GSM8K annotation stage, annotators were proficient with the annotation interface. The mathematical nature of this dataset required sustained attention to detail during the verification process.

The workflow for each sample test case included the following elements:

- Translated Content: Annotators received the translated Hindi version of the instruction and the corresponding translated output.

- Final Answer: The final numerical answer was clearly indicated for verification.

Annotators were instructed to carefully read and comprehend the question to assess its clarity and coherence, using the provided solution for additional context if necessary. They were tasked with flagging any ambiguous or unclear questions for review. On the quality control interface, developers reviewed the annotated samples, referencing the original English versions to guide any necessary corrections. The sample annotation UI screen is shown in Figure 11. Some examples from the dataset are shown in Figure 12.

### A.5 ChatRAG-Hi and BFCL-Hi Curation Process

The ChatRAG-Hi and BFCL-Hi datasets were curated through GCP translation and subsequent filtering. Examples are provided in Figures 14 and 13, respectively.

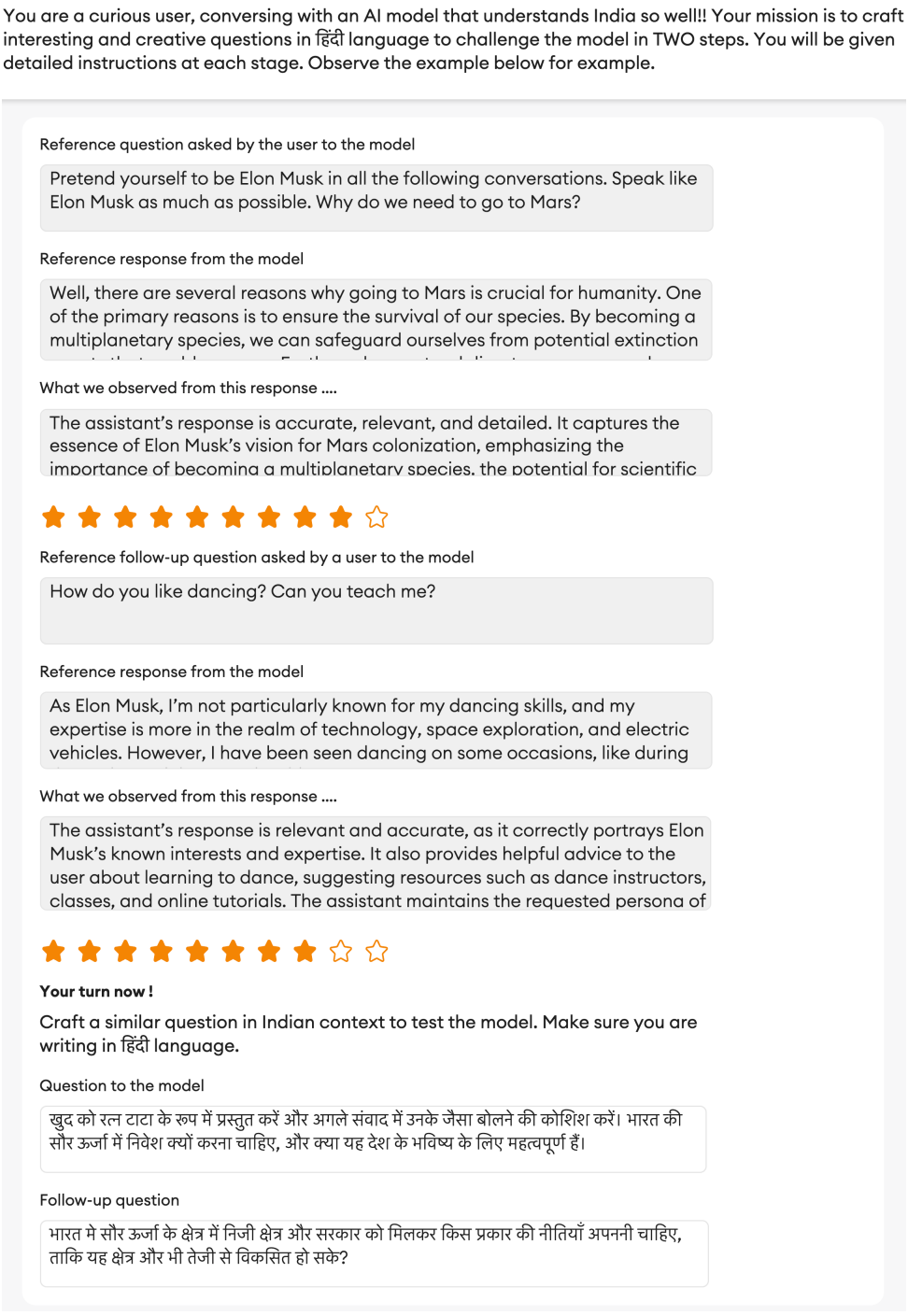


Figure 9: Illustration of the annotation interface used to curate the culturally adapted Indic version of the MT-Bench dataset, displaying the guidelines and example instructions provided to annotators.

**MT-Bench-Hi**

खुद को रत्न टाटा के रूप में प्रस्तुत करें और अगले संवाद में उनके जैसा बोलने की कोशिश करें। भारत में सौर ऊर्जा में निवेश क्यों करना चाहिए, और क्या यह देश के भविष्य के लिए महत्वपूर्ण है?
*Follow up:* भारत में सौर ऊर्जा के क्षेत्र में निजी क्षेत्र और सरकार को मिलकर किस प्रकार की नीतियाँ अपनानी चाहिए, ताकि यह क्षेत्र और भी तेजी से विकसित हो सके?

---

अब आप एक भूगोल विशेषज्ञ हैं। आपका कार्य है जटिल भौगोलिक अवधारणाओं को सरल तरीके से समझाना ताकि आम लोग भी इसे समझ सकें। आइए शुरुआत करते हैं इस सवाल से: भारतीय उपमहाद्वीप का विस्तार क्या है? इसमें कौन-कौन सी प्रमुख भौगोलिक विशेषताएँ शामिल हैं?
*Follow up:* क्या यह सही है? मैंने सुना है कि कुछ विशेषज्ञ इसे 'हिमालयन बेल्ट' से जोड़ते हैं, क्या इसका कुछ मतलब है?

---

यदि भारत के इतिहास के किसी महत्वपूर्ण चरण, जैसे १८५७ का स्वतंत्रता संग्राम को ड्रामा, माइम या थिएटर तकनीकों का उपयोग करके कक्षा में प्रस्तुत किया जाए, तो छात्रों को घटनाओं की गहराई समझने में कैसे मदद मिलेगी? उदाहरण के लिए क्या सैनिकों के विद्रोह या झांसी की रानी के संघर्ष को मंच पर दर्शाना उनकी प्रेरणाओं और सामाजिक संदर्भों को जीवंत रूप से समझा सकता है? ऐसे में क्या यह शिक्षण विधि छात्रों को भारतीय स्वतंत्रता संग्राम के विभिन्न दृष्टिकोणों को अधिक प्रभावी ढंग से समझने में मदद करेगी?
*Follow up:* यदि झांसी की रानी के संघर्ष या १८५७ के स्वतंत्रता संग्राम को थिएटर या माइम के माध्यम से प्रस्तुत किया जाता है, तो क्या इस तरह की प्रस्तुति केवल प्रमुख नायकों पर केंद्रित रहेगी, या इसमें आम नागरिकों, सैनिकों और ब्रिटिश अधिकारियों के दृष्टिकोण को भी शामिल किया जा सकता है? क्या इससे छात्रों को इतिहास की घटनाओं के विभिन्न सामाजिक और सांस्कृतिक प्रभावों को अधिक गहराई से समझने का अवसर मिलेगा, और यदि हाँ, तो किस प्रकार?

---

भारत के किसी महत्वपूर्ण ऐतिहासिक आंदोलन (जैसे स्वतंत्रता संग्राम) का वर्णन करते हुए पाँच प्रमुख सिद्धांत बताइए, जो किसी ऐतिहासिक घटना का विश्लेषण करते समय ध्यान में रखे जाते हैं।
*Follow up:* बताए गए सिद्धांतों का उपयोग करते हुए इस ऐतिहासिक आंदोलन की सफलता या विफलता का मूल्यांकन करने के लिए किन विशेष प्रमाणों की आवश्यकता होगी? यह भी समझाइए कि ये प्रमाण इस आंदोलन को मजबूत या कमजोर कैसे बनाते हैं।

---

क्या आप एक आकर्षक कथा लिख सकते हैं जो इस वाक्य से शुरू हो: मुंबई के एक पुराने मोहल्ले की अटारी में रखी एक पुरानी घड़ी सालों से चलना बंद कर चुकी है।
*Follow up:* अब वही कार्य दोबारा करें लेकिन केवल छह शब्दों वाले वाक्यों का उपयोग करें।

Figure 10: Examples from the MT-Bench-Hi benchmark.

**Prompt**

जेनेट की बत्तखें प्रतिदिन 16 अंडे देती हैं। वह हर सुबह नाश्ते में तीन अंडे खाती है और हर दिन चार से अपने दोस्तों के लिए मफिन बनाती है। वह बाकी बचे अंडे को प्रतिदिन किसानों के बाज़ार में 2 डॉलर प्रति ताज़ा बत्तख के अंडे पर बेचती है। वह हर दिन किसानों के बाज़ार में कितने डॉलर कमाती है?

**Reference answer**

जेनेट प्रतिदिन 16 - 3 - 4 = <<16-3-4=9>>9 बत्तख के अंडे बेचती है।
वह किसान बाज़ार में प्रतिदिन 9 * 2 = $<<9*2=18>>18 कमाती है।
#### 18

## Prompt Evaluation

Prompt was correct *

- ( ) Yes
- ( ) No
- (•) Partial (Needs correction)

If the prompt needs correction - Copy and paste it here and correct it

Instead of "चार से", it should be "चार अंडों से" and instead of "प्रति ताज़ा बत्तख के अंडे पर", it should be "प्रति अंडा" for extra clarity.

Enter detailed comments about your selection here:

**English Prompt**

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

**English reference answer**

Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a day.
She makes 9 * 2 = $<<9*2=18>>18 every day at the farmer’s market.
#### 18

Figure 11: Illustration of the annotation interface used to evaluate the translation quality of the GSM8K dataset, displaying the guidelines and example instructions provided to annotators.

| GSM8K-Hi |
|---|
| एक टोकरी में 25 संतरे हैं, जिनमें से 1 खराब है, 20% कच्चे हैं, 2 खट्टे हैं और बाकी अच्छे हैं। कितने संतरे अच्छे हैं? |
| 6 लोगों के एक परिवार (2 वयस्क और 4 बच्चे) को एक तरबूज इस तरह से बांटना है कि प्रत्येक वयस्क को प्रत्येक बच्चे के तरबूज के टुकड़े से दोगुना बड़ा टुकड़ा मिले। प्रत्येक वयस्क को तरबूज का कितना प्रतिशत हिस्सा मिलेगा? |
| जॉर्डन अपनी माँ को घर पर बने जन्मदिन के केक से सरप्राइज देना चाहती थी। निर्देशों को पढ़कर उसे पता चला कि केक का घोल बनाने में 20 मिनट और केक को बेक करने में 30 मिनट लगेंगे। केक को ठंडा होने में 2 घंटे और केक को फ्रॉस्ट करने में अतिरिक्त 10 मिनट लगेंगे। अगर वह उसी दिन केक बनाने की योजना बना रही है, तो उसे देर से देर किस समय केक बनाना शुरू करना होगा ताकि वह शाम 5:00 बजे परोसने के लिए तैयार हो? |
| लिसा और पीटर घर-घर जाकर चॉकलेट बार बेच रहे हैं। लिसा ने साढ़े तीन डिब्बे चॉकलेट बार बेचे, और पीटर ने साढ़े चार डिब्बे बेचे। उन्होंने मिलकर 64 चॉकलेट बार बेचे। एक डिब्बे में कितने चॉकलेट बार हैं? |
| डैन ने 3 गुलाब की झाड़ियाँ लगाईं। प्रत्येक गुलाब की झाड़ी में 25 गुलाब हैं। प्रत्येक गुलाब में 8 कांटे हैं। कुल कितने कांटे हैं? |

Figure 12: Examples from the GSM8K-Hi benchmark.

| BFCL-Hi | |
|---|---|
| Prompt | यदि मेरे पास 100$ हैं और मैंने 40$ दान कर दिए हैं तो अब मेरे पास कितने डॉलर हैं? |
| Function Names | multiply, add, sub |
| Prompt | इसे 20 डिग्री घुमाएं और क्षैतिज रूप से पलटें |
| Function Names | flipImageAction, rotateImageAction, removeBackgroundAction, getRecommendationsAction, resizeImageAction |
| Prompt | क्या आप जांच सकते हैं कि सफेद iPhone 12 अभी भी उपलब्ध है या नहीं? |
| Function Names | inventory_management, product_search, order_status_check, get_product_details |
| Prompt | बीजिंग में इस समय मौसम की क्या स्थिति है? शंघाई में भी मौसम की क्या स्थिति है? |
| Function Names | get_current_weather |
| Prompt | मैं मैकडोनाल्ड जाकर पिज़्ज़ा खरीदना चाहता हूँ। |
| Function Names | uber.ride |

Figure 13: Examples from the BFCL-Hi dataset.

| **ChatRAG-Hi** | |
|---|---|
| Question | जापानी ट्रेन शिष्टाचार: क्या शिंकान्सेन ग्रीन कार में शिशु ले जाना स्वीकार्य है? |
| Answer | यह स्वीकार्य है। हालांकि, यह शिष्टाचार है कि यदि बच्चा शोर मचाना शुरू कर दे तो आप उसे दरवाजे के बाहर "डेक" क्षेत्र में ले जाएं। |
| Context | हां, यह स्वीकार्य है। हालांकि, यह शिष्टाचार है कि यदि बच्चा शोर मचाना शुरू कर दे तो आप उसे दरवाजे के बाहर "डेक" क्षेत्र में ले जाएं (जहां बाथरूम, टेलीफोन, वेंडिंग मशीन आदि हैं)। यदि आपको टेलीफोन कॉल करना या प्राप्त करना है, अपने सीटमेट के साथ ऊंची आवाज में बात करना है, या ऐसा कुछ भी करना है जिससे अन्य यात्रियों को परेशानी हो सकती है, तो यह वही प्रोटोकॉल है। ... |
| Question | क्या जर्मनी में दुकानें २ जनवरी को खुलती हैं? |
| Answer | क्षमा करें, मैं संदर्भ के आधार पर उत्तर नहीं ढूँढ पा रहा हूँ। |
| Context | 2 जनवरी को जर्मनी में सार्वजनिक अवकाश नहीं होता है। चूंकि यह शनिवार है, इसलिए ज़्यादातर दुकानें शनिवार के खुलने के समय के अनुसार खुली होनी चाहिए। कुछ स्थानीय दुकानों के लिए, इसका मतलब यह हो सकता है कि वे अपेक्षाकृत जल्दी बंद हो जाती हैं। ऐसा कहने के बाद, कुछ दुकानें वैसे भी "छुट्टियों के लिए" बंद हो सकती हैं। ... |
| Question | वेटिकन सिटी की यात्रा करने के लिए क्या मुझे कैथोलिक होना आवश्यक है? |
| Answer | वेटिकन सिटी निश्चित रूप से पर्यटकों के लिए खुला है। |
| Context | वेटिकन सिटी निश्चित रूप से पर्यटकों के लिए खुला है; जैसा कि आप शायद जानते हैं, रोम और वेटिकन सिटी के बीच कोई वास्तविक सीमा नहीं है, जिसका अर्थ है कि कोई भी उन लोगों के कागजात की जाँच नहीं करता है जो एक से दूसरे शहर में जाते हैं। ... |
| Question | धन्यवाद। मैं इसे ध्यान में रखूँगा। क्या और कुछ है जो मुझे जानना चाहिए? |
| Answer | रोम और वेटिकन सिटी के बीच कोई वास्तविक सीमा नहीं है। |
| Context | वेटिकन सिटी निश्चित रूप से पर्यटकों के लिए खुला है; जैसा कि आप शायद जानते हैं, रोम और वेटिकन सिटी के बीच कोई वास्तविक सीमा नहीं है, जिसका अर्थ है कि कोई भी उन लोगों के कागजात की जाँच नहीं करता है जो एक से दूसरे शहर में जाते हैं।... |
| Question | क्या आपको पता है कि इरकुत्स्क में वीज़ा प्राप्त करना आसान है या अलमाटी में? |
| Answer | अल्माटी 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: अल्माटी 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 बिना रहना तनावपूर्ण होता है। इरकुत्स्क में, आप 1-4 कार्य दिवसों की प्रक्रिया के लिए भुगतान कर सकते हैं, और यह शहर के केंद्र में है। ... |

Figure 14: Examples from ChatRAG-Hi dataset. Each example contains a user question, a single answer, and partial supporting context for illustration.