BharatBench: Comprehensive Multilingual Multimodal Evaluations of Foundation AI models for Indian Languages

Krutrim AI Team [1]

Abstract

This paper presents a comprehensive evaluation of Foundation AI models in multilingual and multi-modal contexts, addressing their performance, adaptability, and limitations. With the increasing globalization of digital content and the rise of multi-modal data, understanding how Large Language Models (LLMs) operate across diverse languages and modes of information is critical. We introduce *BharatBench* as an evaluation framework benchmarking the performance of multilingual LLMs. We conduct a series of experiments using benchmark datasets that encompass various languages and modalities, including text, images, and audio. We assess a diverse array of proprietary and open-source language models across this benchmark. Our findings underscore the need for improved training strategies and dataset curation to enhance the efficacy of LLMs in real-world multilingual and multi-modal applications. We maintain the leaderboard of BharatBench at `https://cloud.olakrutrim.com/console/inference-service?section=leaderboards.ri`

1. Introduction

Recent advancements in Large Language Models (LLMs) (Brown et al., 2020; Google et al., 2023; Achiam et al., 2023; Gemma et al., 2024; Jiang et al., 2024; Touvron et al., 2023a;b; Dubey et al., 2024) have significantly enhanced performance, demonstrating advanced capabilities in complex language tasks such as question-answering, summarization, machine translation, etc. as well as perception and speech related tasks. Despite these advancements, evaluating LLM performance across diverse languages, particularly those less represented than English, Chinese or European languages, presents substantial challenges. Recent studies highlight a performance gap between high-resource and low-resource languages in text generation capabilities (Ahuja et al., 2023a;b).

NLP practitioners looking for Artificial Intelligence (AI) models tailored to Indian languages and regional use cases often encounter difficulties due to a lack of comprehensive information and evaluation tools. This often leads them to select generic, non-optimized models, resulting in suboptimal performance, especially for India's native multilingual population of over 1 billion citizens.

To this end, we propose an evaluation framework *BharatBench*, which addresses this critical pain point by providing a clear, data-driven leaderboard that evaluates AI models and products based on their performance with Indian languages and use cases. BharatBench leverages proprietary and public datasets that encompass the linguistic and cultural diversity of the Indic region, ensuring that evaluations are highly relevant and representative of real-world use cases in India. This makes BharatBench a critical tool for organizations and researchers looking to optimize their AI models and solutions for the Indian market while promoting greater inclusivity in the AI ecosystem.

We propose a comprehensive methodology that leverages diverse benchmark datasets to assess model performance across various languages and modalities including speech and perception. Compared to prior works, we are the first to provide multimodal benchmark encompassing perception and speech as well as decompose the text generation capabilites for morphologically rich low-resource languages into sub-tasks, which we expand upon in Section 3.1.

Our contributions could be summarized as:

- We create and release BharatBench, the first comprehensive evaluation framework, that benchmarks text generation capabilities as well as visual and speech understanding in the multi-lingual setting.
- We cover cultural and language diversity as part of BharatBench, supporting 8 Indic languages and include evaluation across multiple cultural artifacts.
- We evaluate different SOTA models across the benchmark, establishing respective baselines across different

[1] Krutrim AI, Bangalore, India. Correspondence to: Chandra Khatri <chandra.khatri@olakrutrim.com>.

modalities. Our leaderboard would be made publicly available.

The rest of the paper is organized as follows: Section 2 presents related research on multilingual evaluation while Section 3 explains our evaluation framework in detail. We present the experimental findings in Section 4, followed by conclusions in Section 5.

2. Related Work

Limited number of LLMs have been specifically designed for Indic languages (Labs, 2023; Balachandran, 2023; Kohli et al., 2023; Gala et al., 2024; Sarvam, 2023; Choudhury et al., 2024), primarily extending and fine-tuning text-only English-centric models with an exception of models like (Krutrim, 2024; Bendale et al., 2024; Sarvam, 2024), trained from scratch. In parallel, there have been efforts to create multi-linugal datasets, specifically targeting Indian languages. IndicNLP corpora (Kunchukuttan et al., 2020) was created by scraping web sources - primarily news, magazines and books (2.7B tokens across 10 Indian languages). It was extended to form IndicCorp (Kakwani et al., 2020), 8.8B tokens across 11 major Indian languages and English. Indic Instruct Data v0.1 was introduced by (Gala et al., 2024) as Hindi instruction tuning dataset created by translating existing sources. Sangraha (Khan et al., 2024a) is of higher magnitude containing 251B tokens over 22 languages, as a large scale pretraining data.

Most existing research has, however, primarily focused on LLM performance in monolingual settings, often overlooking the challenges presented by linguistic diversity and underrepresented languages. In the context of LMMs, few recent works explore India-centric language evaluations. IndicGLUE (Indic General Language Understanding Evaluation Benchmark) (Kakwani et al., 2020) is a collection of Natural Language Understanding (NLU) benchmark for Article Genre Classification, Headline Prediction, Wikipedia Section-Title Prediction, Cloze-style Multiple choice QA, Winograd NLI and COPA. IndicXTREME (Doddapaneni et al., 2023) included 9 tasks across sentence classification, structure prediction, question answering and sentence retrieval. Specifically, they collated already existing benchmarks such as IndicXNLI (Aggarwal et al., 2022) -automatic translation of XNLI corpus (Conneau et al., 2018) in 11 languages, MASSIVE (FitzGerald et al., 2022) - intent classification task for 7 languages, Naamapadam (Mhaske et al., 2022) - NER for 9 languages , FLORES-101 (NLLB-Team et al., 2022) and new NLU benchmarks like IndicCOPA, IndicQA, IndicXParaphrase and IndicSentiment

On a similar note, there have been efforts for evaluating the Natural Language Generation (NLG) capabilites of Indic LLMs. IndicNLG Suite (Kumar et al., 2022), consists of 5 NLG taks in 11 Indic languages – biography generation (BG) using Wikipedia infoboxes (WikiBio), news headline generation (HG), sentence summarization (SS), paraphrase generation (PG) and question generation (QG). Most recently, IndicGenBench (Singh et al., 2024b) was introduced comprising tasks like cross-lingual summarization (CROSSSUM-IN), machine translation (FLORES-IN), cross-lingual reading comprehension (XORQA-IN-XX and XORQA-IN-EN) and multilingual reading comprehension (XQUAD-IN). It extends existing benchmarks such as CrossSum (Bhattacharjee et al., 2023), XQuAD (Artetxe et al., 2020), XorQA (Asai et al., 2021), and FLORES (NLLB-Team et al., 2022) for additional Indic languages. Another work, IndicQA benchmark (Singh et al., 2024a) evaluates extractive and abstractive closed question-answering capabilities of LLMs for 11 major Indian languages while Rohera et al. (2024) focuses on evaluating open domain QA without contextual passages for 19 Indian languages.

Recent works also show that LLMs like GPT-4 (Hada et al., 2023) as an evaluator in multilingual settings provide inconsistent biased results for low resource languages, underscoring the need for custom multi-lingual LLM evaluators. FBI (Doddapaneni et al., 2024b) is a novel meta-evaluation framework that assesses the robustness of evaluator LLMs across various tasks and strategies. It highlights significant shortcomings in these models, revealing that they fail to detect quality drops in over 50% of cases, thus emphasizing their unreliable nature and the need for cautious implementation in practical applications Most recently, Doddapaneni et al. (2024a) proposed Cross Lingual Auto Evaluation (CIA) Suite for multilingual evaluation using LLMs, similar to the works of (Li et al., 2023a; Zheng et al., 2023; Kim et al., 2023; 2024; Dubois et al., 2024)

LMSys Chatbot arena[1] (Chiang et al., 2024) is an open-source platform for evaluating AI through human preferences, following pairwise comparison using Elo ratings (Elo & Sloan, 1978). Following the same principles, Pariksha [2] and Health Pariksha (Gumma et al., 2024) were introduced for Indic languages in general as well as medical domain respectively.

Concurrent to our work, (Thellmann et al., 2024) explores the evaluation framework for European languages. In contrast, our work focus on creating evaluation framework across different modalities for specific tasks in Indic languages.

[1] https://lmarena.ai/

[2] https://coda.io/@peopleplusai/glocal-evaluation-of-models

3. BharatBench

BharatBench is an India-centric evaluation framework, designed to rank and evaluate AI models based on their performance on a diverse set of use cases, enabling businesses, researchers, and developers to make informed decisions about model selection and deployment. In the following, we outline the criteria used for different modalities in our evaluation suite.

3.1. Language tasks

As part of our evaluation suite, we compare the performance of LLMs on the text generation and understanding capabilities as well as the distributional representations.

3.1.1. Text Generation and Understanding

We create a dataset of 300 examples across 8 languages and 5 tasks rated by native speakers. Our task selection process ensure diversity in evaluation. This dataset is designed to assess the model's ability to perform across 5 different language-related tasks in 8 languages, ensuring comprehensive coverage of linguistic, cultural, and contextual challenges. Prompts covered a variety of linguistic and contextual challenges, ensuring depth and breadth in evaluation while responses reflect natural language, keeping grammar and clarity intact. We aim for a diverse representation of topics and linguistic features across prompts and responses described below:

1). Indian Cultural Context (ICC) : The Indian Cultural Context (ICC) task refers to the rich and diverse tapestry of customs, traditions, social practices, languages, arts, religions, and historical developments that form the essence of India's cultural identity. This category reflects the complex interplay of regional, religious, linguistic, and social factors that have shaped India's heritage and continues to influence its contemporary social fabric. Cultural relevance and context is required to be the core of such questions thereby encouraging thoughtful and in-depth responses. We ensured that the prompts remain neutral so that they are not restricted to only one region. This is to encourage a more inclusive and comprehensive representation of India's cultural landscape and nuances while also avoiding stereotypes. We ensured that all ground truth responses are written by human contributors. Machine-generated responses are not permitted, as human input is essential to maintain the quality and authenticity of the content. Some examples include: **1).** What is the cultural importance of performing Giddha? **2).** What are the traditional spices used in Chettinad cuisine of Tamil Nadu?

2). Multi-turn comprehension: Comprehension is the ability to understand and make sense of what we read or hear. In reading, it means understanding the meaning of the text, finding important points, and seeing how ideas connect. Comprehension is an important skill because it helps us better understand and remember what we read, allowing us to learn more effectively. The objective of creating a prompt dataset and their responses is to evaluate the model's comprehension abilities. This involves generating diverse prompts to assess the model's understanding and interpretative skills. The dataset covers a range of topics and complexity levels to ensure comprehensive testing. Responses will be analyzed for accuracy, relevance, and depth of understanding. The ultimate goal is to gauge the model's proficiency in interpreting and responding to varied prompts. We use different domain paragraphs in each prompt from Movies, Sports, Social Science, Health, Technology, Politics, and additional relevant domains. Given a comprehension passage, example questions can be: **1).** Who is the main character? **2).** According to the above passage, who was killed by the big heavy branch?

3). Multi-turn Translation: Translation task refers to the automated process of converting text from one language (the source language) to another (the target language), while preserving the original meaning, context, and tone. This task is designed to evaluate a language model's capabilities in translation, specifically focusing on accuracy, fluency, cultural understanding, and other essential factors. The main aim is to create prompts that evaluate the translation performance of LLM across different languages and contexts. To make sure the models are tested on a broad range of linguistic and cultural challenges, each prompt incorporate elements such as Idioms and Colloquial Phrases, Polysemy (Words with Multiple Meanings) as well as acronyms and industry-specific terms.

4). Text classification: The objective of this task is to ensure consistency and clarity in prompt-response creation by defining the structure and expectations for each prompt and response. It also outlines how to provide accurate, relevant, and contextually enriched answers. We measure the efficacy of different models across **1).** Sentiment Analysis **2).** Topic Classification, e.g. Legal, Sports, Technology, Entertainment, etc. **3).** Intent Classification - determining the underlying goal or purpose of a user's utterance such as Request, Command, Feedback, Complaint **4).** Language Identification **5).** Spam Detection

5). Grammar Correction: Grammar correction is a task that involves identifying and correcting grammatical errors in sentences. The objective of this task is to assess the model's ability to accurately understand and correct various grammatical mistakes in different Indic languages including English. An effective dataset to test an LLM's grammar correction capability should cover a wide range of error types, complexities, domains, and prompt lengths to ensure a comprehensive assessment. We examine across the following axes:

- Word Order Errors: Errors where words are misplaced in a sentence, leading to awkward or incorrect sentence structures. Example: He eats quickly the cake.
- Sentence Fragment Errors: Incomplete sentences that lack essential components such as a subject, verb, or complete thought. Example: Went to the market.
- Spelling Mistakes: Errors related to incorrect spelling, including typos, homophones. Example: Word pronounciation means the style of pronouncing words.
- Punctuation Errors: Errors involving missing or misused punctuation marks that affect sentence meaning. Example: Lets eat Grandma.
- Agreement Errors: Errors that occur when the subject and verb do not match in number, gender, or person or when there is lack of agreement between an adjective and the noun. In Indic languages, verbs often agree with the subject in person, gender, and number. We further subdivide these errors as:
    - Number Agreement (Singular/Plural): Errors occur when the verb form does not match the number of the subject. Example: The dogs barks at strangers.
    - Person Agreement (First/Second/Third): This subcategory involves the agreement between the subject and verb based on person (first, second, or third person). Errors occur when the verb form does not match the grammatical person of the subject. Example: He go to the market
    - Gender Agreement (Masculine/Feminine): Errors in gender agreement occur when the verb form does not match the subject's gender.
    - Honorific Agreement (Low/Medium/High): In some Indic languages, verbs must agree with the level of formality or respect implied by the pronoun used. Errors in honorific agreement occur when the verb form does not match the honorific level of the subject or pronoun.
    - Case Agreement: Errors occur when postpositions or case markers do not correctly align with the verb or object. This can lead to confusion about the roles of the subjects, objects, or other sentence elements.
    - Adjective-Noun Agreement: Errors occur when an adjective do not correctly align with the noun it describes. The gender or number of the noun aligns with the adjective it describes in many Indic languages.

We created three types of prompts across all the sub-tasks:

- Identify the Mistake: The LLM is presented with a sentence and asked to identify the grammatical error(s) without necessarily providing the correction. Example: Identify the mistake in this sentence "I will eat whether I feel hungry."
- Correct the Mistake: The LLM is given a sentence containing an error and is asked to provide the correct form of the sentence. Example: Rectify the mistake in this sentence "I will eat whether I feel hungry."
- Fill in the Blanks: LLM is presented with a sentence that has one or more blanks, which has to be filled with correct words to form a grammatically accurate sentence. Example: Despite _ (try) her best, she could not convince her parents to let her go on the trip

3.1.2. Distributional Representations (Embedding)

Embedding models are used to encode sentences, paragraphs and documents into feature representations, which are further used for classification, retrieval etc based on similarity. We propose to develop an embedding model, capable in 9 Indic Languages, including English. The motivation for creating such a model stems from developing general purpose embedding models which can handle Indian languages and context as well as capable of handling large documents. For this use case, we choose sentence retrieval from cross-lingual XTREME benchmark, evaluating on accuracy. We expand upon the model performance in Section 4.2.2

3.2. Visual understanding (Perception)

For this multimodal task, the aim is to create a robust and comprehensive evaluation framework tailored to Indian cultural, social, and environmental specifics, thereby advancing the capabilities and accuracy of Vision Language Models (VLMs) in understanding and interpreting Indian-related visual content. We start by collecting and curating a diverse set of images and questions representing Indian contexts. Some topics that we include are: Bollywood, Livelihood, Seasons, Sports, Art forms, Historical Sites, Dance forms, Wildlife, Weddings, Traditions, Monuments, Cuisines, Festivals, Indian houses, Ancient Rulers, Cityscapes, Village, Landscape, Actors, People, to name a few. Our evaluation framework consists of 3 aspects:

3.2.1. Translated English academic datasets

We first include the translated versions of popular English academic datasets using IndicTransV2 (Gala et al., 2023). This is done to assess general multi-modal capabilites of the VLMs in 8 different languages. Specifically, our multilingual evaluation framework includes translation of:

- POPE (Li et al., 2023b): Assess hallucination tendencies - Yes/No responses about visual objects in images
- LLaVA-Bench (In-the-Wild) (Liu et al., 2023): Visual question answering in uncontrolled, real-world environments.
- GQA (Hudson & Manning, 2019): Visual perception, i.e. questions around multiple reasoning skills from the real world images, spatial scene understanding and multi-step inference.
- MM-Vet (Yu et al., 2023): Recognition, OCR, knowledge, language generation, spatial awareness, and math.

3.2.2. BharatBench-V Dataset

The BharatBench-Vision dataset was curated from regional and national newspaper websites, comprising 30 images per language for a total of 9 Indian languages, including English. Each image was associated with 5 prompts, covering Visual Question and Answering, Classification, Yes/No Questions, Caption Generation, and Adversarial Question and Answering tasks. To ensure cultural relevance, a team of annotators and linguists selected 30 representative images from an initial pool of around 1,000 images per language. These images were chosen based on predefined categories such as Indian festivals, cultural dances, Bollywood, etc. Subsequently, in-house language experts generated 5 question-answer pairs for each image, ensuring accurate and diverse annotations across all languages.

3.2.3. BharatBench-OCR Dataset

We also propose OCR capabilites on Indian content as part of the visual understanding capabilites. We evaluate different OCR models on approx 40 images from the internet. We ensured that these images were not present in the training set (by not taking images from books present at US archive). This set majorly consists of scanned book pages(in mostly hindi but also in sanskrit and a bit of english and marathi) with various layouts like single column, double column, mixed layout like (two column + one column), images in between the paragraphs, two-page photos, text with image in the background, slightly skewed images (that are at an angle, not rotated, like half open book), homework pages (containing match the following, fill in the blanks), index pages, clean handwritten text, hindi with old characters. We prepared the ground truth from these images using in-house annotators and computed Character Error Rates (CER) and Word Error Rates (WER) model outputs against the ground truth.

3.3. Speech Understanding

Our evaluation framework for speech modality supports 8 Indic languages - Hindi, Marathi, Tamil, Telugu, Kannada, Gujarati, Bengali, Malayalam. We first collected approximately 800 hours of proprietary data across the 8 languages (100 for each language) by 50 native speakers in total. We also used proprietary translation service to translate the transcripts of the dataset into different languages and evaluate across the following two tasks:

- **Speech to Text Transcription:** For the subjective evaluation for this task, we sample 300 examples for each language, which had diversity across multiple domains. Annotators were asked to mark the transcription as good or bad, depending on the meaning of the transcript being identical to the reference text spoken in the audio.
- **Speech to Text Translation:** Similar to the previous task, we sample 300 examples for each language-pair, which is diverse in domains. Annotators were asked to mark the translation as good or bad, depending on the meaning of translation being identical to the input text.

4. Experimental Results

In this section, we provide baseline results for different models on the BharatBench discussed above.

4.1. Tokenization

We built a custom tokenizer, specifically tailored for Indian languages. Tokenizer plays an important role when effective sequence length is of significance. However, if a tokenizer splits individual words into higher number of tokens, this will result in lesser effective sequence length. We provide a comparative analysis of different proprietary and open-source models along with LLMs trained from scratch on Indian languages. Specifically, we consider GPT-4, GPT-4o, LLaMA3-8B, Gemma-2B (Gemma et al., 2024), Nemotron4 (Adler et al., 2024; Parmar et al., 2024), Minitron (Sreenivas et al., 2024), Mistral-Nemo as well as Indic-specific LLMs such as Sarvam-2B, Sutra and Krutrim LLM. Table 1 shows fertility rate i.e. token-to-word ration of different tokenizers for Indian languages, where lower fertility rate is desirable. Krutrim tokenizer achieves lower fertility rate for 10 out of 11 Indian languages while having reasonable performance for English as well as code data.

4.2. Language Tasks

We provide a detailed analysis of different models for language tasks discussed in Section 3.1

Model	Assamese	Bengali	Bodo	Dogri	English	Konkani	Gujarati	Hindi	Kashmiri	Kannada	Maithili	Malayalam	Manipuri	Marathi	Nepali	Odia	Punjabi	Sanskrit	Santali	Sindhi	Tamil	Telugu	Urdu
gpt4 (100K)	8.52	8.43	6.50	4.38	1.36	5.43	9.78	5.30	4.01	14.28	4.87	16.32	5.86	7.10	6.54	16.06	8.03	8.88	12.22	3.97	12.30	13.70	3.96
gpt-4o(200K)	2.77	2.51	3.16	1.89	1.34	2.72	2.38	1.77	1.59	3.35	2.02	3.69	2.41	2.59	2.11	6.42	2.69	3.94	13.01	1.75	3.28	3.39	1.51
Llama-3-8B(128256)	8.44	8.35	3.63	2.92	1.36	3.44	9.78	2.80	2.73	14.28	2.79	16.32	5.31	3.92	3.52	15.72	7.96	4.83	12.17	2.87	12.30	13.69	2.75
gemma2b (256K)	4.34	3.89	3.47	2.17	1.39	3.12	3.90	2.12	1.85	5.51	2.44	6.07	3.02	3.23	2.88	6.38	3.34	4.21	5.11	2.34	4.32	4.86	1.75
Nemotron4-340Base (256K)	4.84	2.78	3.35	2.04	1.39	2.95	15.47	1.93	2.45	4.37	2.30	4.90	2.65	2.74	2.34	16.98	12.81	4.39	13.07	2.69	3.73	4.05	1.59
Mistral-Nemo	4.41	2.89	3.52	2.12	1.41	3.08	3.68	2.07	1.82	3.87	2.48	4.88	2.67	3.13	2.97	16.95	3.10	4.34	12.16	2.51	3.69	3.73	1.65
Sarvam-2B	4.40	2.00	2.92	1.85	1.68	3.01	2.17	1.55	1.91	2.60	2.11	3.31	4.60	1.97	2.35	2.47	1.76	3.78	13.07	7.62	2.53	2.68	7.93
Sutra	2.18	2.11	3.06	1.78	1.18	2.68	2.18	1.64	1.48	2.73	2.08	3.15	2.40	2.20	2.01	2.27	1.52	3.76	2.03	2.23	2.60	2.78	1.55
Krutrim - 200K	1.90	1.86	1.82	1.58	1.42	2.18	1.84	1.36	1.38	2.15	1.57	2.43	2.33	1.61	1.53	1.82	1.59	2.53	1.58	1.58	2.07	2.10	1.52

Table 1. **Performance of tokenizer across different languages where a lower fertility rate is desirable.** We provide a comparative analysis of different proprietary and open-source models along with LLMs trained from scratch on Indian languages.

Bench	Bengali	English	Gujarati	Hindi	Kannada	Malayalam	Marathi	Tamil	Telugu
Indian Cultural Context (BERT Score (0-shot))									
Llama-3.2-3B-Instruct-Turbo	0.82	0.89	0.87	0.88	0.85	0.85	0.83	0.86	0.86
Meta-Llama-3.1-70B-Instruct-Turbo	0.84	0.9	0.89	0.89	0.86	0.88	0.87	0.89	0.88
Meta-Llama-3.1-8B-Instruct-Turbo	0.83	0.89	0.87	0.88	0.85	0.87	0.85	0.88	0.87
Mistral-Nemo-Instruct-2407	0.8	0.89	0.84	0.85	0.83	0.83	0.82	0.84	0.85
gemma-2-27b-it	0.83	0.89	0.88	0.88	0.86	0.87	0.85	0.88	0.88
gemma-2-9b-it	0.83	0.9	0.88	0.88	0.85	0.86	0.85	0.88	0.88
gpt-4o	0.85	0.9	0.89	0.9	0.87	0.88	0.88	0.91	0.89
gpt-4o-mini	0.85	0.91	0.89	0.89	0.86	0.88	0.87	0.9	0.89
Krutrim-1	0.85	0.9	0.9	0.89	0.87	0.89	0.87	0.9	0.89
Krutrim-2	0.85	0.86	0.90	0.89	0.87	0.88	0.88	0.90	0.89
Multi-Turn Comprehension (BERT Score (0-shot))									
Llama-3.2-3B-Instruct-Turbo	0.91	0.93	0.87	0.92	0.92	0.89	0.91	0.9	0.92
Meta-Llama-3.1-70B-Instruct-Turbo	0.92	0.94	0.87	0.93	0.94	0.93	0.93	0.93	0.94
Meta-Llama-3.1-8B-Instruct-Turbo	0.91	0.94	0.86	0.91	0.93	0.92	0.92	0.92	0.93
Mistral-Nemo-Instruct-2407	0.88	0.93	0.83	0.89	0.91	0.89	0.88	0.87	0.9
gemma-2-27b-it	0.92	0.93	0.84	0.91	0.94	0.94	0.92	0.91	0.92
gemma-2-9b-it	0.91	0.93	0.85	0.91	0.93	0.91	0.92	0.89	0.93
gpt-4o	0.94	0.96	0.85	0.95	0.96	0.94	0.94	0.93	0.96
gpt-4o-mini	0.94	0.96	0.85	0.94	0.95	0.92	0.93	0.91	0.94
Krutrim-1	0.93	0.94	0.89	0.93	0.95	0.91	0.94	0.92	0.93
Krutrim-2	0.93	0.94	0.88	0.94	0.92	0.92	0.94	0.91	0.90
Multi-Turn Translation (BERT Score (0-shot))									
Llama-3.2-3B-Instruct-Turbo	0.88	0.88	0.86	0.88	0.85	0.84	0.87	0.87	0.89
Meta-Llama-3.1-70B-Instruct-Turbo	0.9	0.9	0.93	0.9	0.91	0.91	0.9	0.91	0.94
Meta-Llama-3.1-8B-Instruct-Turbo	0.89	0.89	0.89	0.9	0.89	0.88	0.88	0.89	0.9
Mistral-Nemo-Instruct-2407	0.86	0.88	0.9	0.87	0.86	0.87	0.86	0.87	0.9
gemma-2-27b-it	0.88	0.88	0.94	0.89	0.92	0.91	0.89	0.92	0.93
gemma-2-9b-it	0.88	0.89	0.92	0.89	0.9	0.9	0.89	0.91	0.93
gpt-4o	0.91	0.91	0.95	0.9	0.92	0.9	0.9	0.92	0.95
gpt-4o-mini	0.91	0.91	0.94	0.91	0.92	0.9	0.91	0.92	0.95
Krutrim-1	0.89	0.9	0.92	0.89	0.91	0.89	0.88	0.91	0.91
Krutrim-2	0.91	0.90	0.94	0.89	0.93	0.93	0.91	0.93	0.94
Multi-Turn (BERT Score (0-shot))									
Llama-3.2-3B-Instruct-Turbo	0.88	0.89	0.88	0.9	0.88	0.85	0.88	0.87	0.86
Meta-Llama-3.1-70B-Instruct-Turbo	0.9	0.91	0.9	0.91	0.9	0.88	0.9	0.9	0.91
Meta-Llama-3.1-8B-Instruct-Turbo	0.89	0.9	0.89	0.9	0.88	0.86	0.89	0.86	0.88
Mistral-Nemo-Instruct-2407	0.88	0.9	0.88	0.89	0.87	0.84	0.87	0.85	0.85
gemma-2-27b-it	0.9	0.9	0.89	0.91	0.9	0.87	0.9	0.87	0.89
gemma-2-9b-it	0.9	0.89	0.89	0.91	0.89	0.86	0.9	0.87	0.87
gpt-4o	0.92	0.92	0.92	0.93	0.92	0.89	0.92	0.9	0.92
gpt-4o-mini	0.92	0.92	0.91	0.92	0.91	0.87	0.91	0.89	0.9
Krutrim-1	0.92	0.92	0.92	0.92	0.91	0.88	0.91	0.89	0.9
Krutrim-2	0.92	0.91	0.92	0.92	0.92	0.88	0.91	0.88	0.90
Grammar Correction (BERT Score (5-shot))									
Llama-3.2-3B-Instruct-Turbo	0.95	0.97	0.96	0.97	0.95	0.94	0.87	0.95	0.96
Meta-Llama-3.1-70B-Instruct-Turbo	0.98	0.99	0.98	0.99	0.97	0.97	0.95	0.98	0.98
Meta-Llama-3.1-8B-Instruct-Turbo	0.96	0.97	0.96	0.98	0.94	0.95	0.87	0.96	0.97
Mistral-Nemo-Instruct-2407	0.95	0.97	0.95	0.98	0.95	0.93	0.85	0.95	0.93
gemma-2-27b-it	0.98	0.98	0.97	0.98	0.97	0.97	0.87	0.97	0.97
gemma-2-9b-it	0.96	0.98	0.97	0.98	0.96	0.97	0.87	0.96	0.98
gpt-4o	0.98	0.98	0.97	0.98	0.97	0.98	0.87	0.98	0.98
gpt-4o-mini	0.98	0.99	0.97	0.99	0.97	0.98	0.87	0.97	0.98
Krutrim-1	0.96	0.98	0.98	0.98	0.97	0.97	0.94	0.97	0.97
Krutrim-2	0.98	0.99	0.98	0.98	0.98	0.98	0.94	0.98	0.98

Table 2. **Performance of different models on BharatBench language generation and understanding tasks.** We report accuracy for Text classification task while BERT Score for the others.

Bench	Bengali	English	Gujarati	Hindi	Kannada	Malayalam	Marathi	Tamil	Telugu
Text Classification (Accuracy (0-shot))									
Llama-3.2-3B-Instruct-Turbo	0.77	0.8	0.33	0.57	0.47	0.48	0.63	0.32	0.58
Meta-Llama-3.1-70B-Instruct-Turbo	0.87	0.92	0.5	0.93	0.8	0.77	0.88	0.75	0.9
Meta-Llama-3.1-8B-Instruct-Turbo	0.8	0.83	0.33	0.77	0.58	0.52	0.73	0.6	0.75
Mistral-Nemo-Instruct-2407	0.83	0.83	0.45	0.88	0.65	0.62	0.73	0.7	0.78
gemma-2-27b-it	0.83	0.92	0.67	0.95	0.72	0.75	0.83	0.73	0.9
gemma-2-9b-it	0.78	0.87	0.6	0.9	0.7	0.68	0.82	0.72	0.8
gpt-4o	0.83	0.93	0.68	0.97	0.85	0.7	0.85	0.8	0.9
gpt-4o-mini	0.8	0.93	0.65	0.93	0.75	0.72	0.88	0.78	0.88
Krutrim-1	0.75	0.77	0.53	0.72	0.65	0.6	0.6	0.62	0.73
Krutrim-2	0.82	0.92	0.50	0.83	0.68	0.65	0.72	0.68	0.78
Named Entity Recognition (Accuracy (5-shot))									
Meta-Llama-3.1-70B-Instruct-Turbo	0.55	0.72	0.82	0.79	0.39	0.49	0.77	0.52	0.46
Meta-Llama-3.1-8B-Instruct-Turbo	0.47	0.69	0.69	0.74	0.35	0.35	0.74	0.46	0.49
Mistral-Nemo-Instruct-2407	0.51	0.67	0.54	0.71	0.31	0.4	0.64	0.49	0.49
gemma-2-27b-it	0.67	0.79	0.75	0.8	0.45	0.49	0.79	0.58	0.55
gemma-2-9b-it	0.56	0.73	0.72	0.76	0.47	0.46	0.73	0.54	0.49
gpt-4o	0.59	0.77	0.79	0.8	0.41	0.52	0.76	0.62	0.56
gpt-4o-mini	0.53	0.76	0.6	0.81	0.4	0.43	0.77	0.55	0.54
Krutrim-1	0.43	0.51	0.21	0.58	0.2	0.29	0.12	0.43	0.24
Krutrim-2	0.51	0.63	0.71	0.67	0.32	0.46	0.73	0.45	0.41

Table 3. **Performance of different models on BharatBench text classification and NER tasks.** We report 0-shot accuracy for Text Classification and 5-shot accuracy for the NER task.

4.2.1. Text Generation and Understanding

Baselines: In this section, we analyze the performance of various models in different Indic languages, as shown in Table 2 and 3 for the language generation and understanding tasks described in Section 3.1.1. We compare Krutrim against Llama-3.1-70B-Instruct-Turbo, Llama-3.1-8B-Instruct-Turbo, Mistral-Nemo-Instruct-2407, Gemma-2-27b-it, Gemma-2-9b-it, GPT-4o and GPT-4o-mini.

Results and Analysis: We report accuracy for Text classification task while BERT Score for the others. For the ICC task, we find Krutrim-2 to be performing best for most of the languages among the considered baselines. GPT-4 performs comparable and ties to Krutrim on the other tasks such as Multi-turn comprehension, translation and Grammar correction tasks. Notably, for the text classification, LLaMA 3.1-70B Instruct also performs reasonably well, compared to Krutrim series and GPT-4o.

4.2.2. Distributional Representations (Embedding)

Baselines: In this section, we analyze the performance of various models across different Indic languages, as shown in Table 4. The models evaluated include MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020) IndicBERT-v2 and a proprietary model *Vyakyarth* (Pandey & Panuganti, 2025).

Results and Analysis: The key findings are discussed below. Overall, the results indicate that the Vyakyarth model significantly outperforms the other models, achieving an average score of 97.8, which is notably higher compared to other embeddings such as Jina-embeddings-v3 (96.0) The high performance of Vyakyarth can be attributed to its improved training approach, specifically tailored for cross-lingual sentence similarity in Indic languages. This model achieves near-perfect performance across multiple languages, demonstrating its robustness and generalizability. This demonstrates that the Vyakyarth model effectively captures the nuances of multiple Indic languages, especially in scenarios where data scarcity is an issue.

Comparing the baseline models, IndicBERT and MuRIL, we find that IndicBERT generally outperforms MuRIL across most languages. IndicBERT achieved an average score of 69.4, compared to MuRIL's 52.3. However, both models struggle with Sanskrit (sa), which highlights the challenges that these models face in low-resource languages. The addition of data augmentation strategies, such as Samanantar and Back-Trans, results in significant performance gains for IndicBERT, increasing the scores across several languages, including Hindi (hi), Marathi (mr), and Tamil (ta).

It is also noteworthy that the Jina embeddings (jina-embeddings-v3) model performs consistently well, achieving an average score of 96.0. The model's competitive performance suggests that the underlying multilingual training approach is effective for a wide range of Indic languages.

In conclusion, the results indicate that while baseline models such as MuRIL and IndicBERT have laid the foundation for multilingual embeddings, the newer models such as Vyakyarth and Jina-embeddings-v3 provide significant advancements in performance. These models demonstrate the importance of incorporating contrastive learning and multi-

Model	Bengali	Gujarati	Hindi	Kannada	Malayalam	Marathi	Tamil	Telugu	Sanskrit	Avg.
MuRIL	77	0	67	84	88	82	36	79	84	67
IndicBERT	91	92	91	89	89	93	90	89	30	84
IndicBERT+Samanantar	89	88	86	88	86	90	88	88	18	80
IndicBERT+Back-Trans	91	91	94	90	88	94	91	89	41	85
IndicBERT-SS	92	86	85	88	92	86	89	87	37	82
XLMR-STSB	90	87	100	87	90	78	86	86	72	86
Jina-embeddings-v3	97	97	99	97	96	97	96	97	84	96
Vyakyarth-v0	96	93	99	93	94	95	96	94	84	94
Vyakyarth-mini	99	99	100	99	99	99	98	98	90	98

Table 4. Performance of Indic Embedding models across different languages

lingual fine-tuning to handle the diversity and complexity of Indic languages effectively.

4.3. Visual understanding

4.3.1. Baselines

We compare the VLM performance of Chitrarth model (Khan et al., 2024b) against LLaMA 3.2-11B-V-Instruct for the BharatBench-V dataset as well as translated multi-modal academic datasets. For the BharatBench-OCR dataset, we consider proprietary GPT-4o and GCP as additional baselines. We trained a custom OCR model Shabdarth specifically for Indian languages. The OCR model was trained on a dataset of approximately 2 million samples, consisting primarily of scanned PDFs of Hindi and Sanskrit texts. A significant challenge arose from the inclusion of old Sanskrit books, which contain many obsolete characters and complex ligatures. Additionally, the dataset featured multi-column Sanskrit-Hindi books, where one column presented Sanskrit shlokas and the other their Hindi translations. The model had to not only accurately recognize intricate characters but also learn the correct reading order for these dual-language, multi-column documents.

4.3.2. Results and Analysis

Figure 1 shows the Character Error Rate (CER) and Word Error Rate of the Shabdarth model against GPT-4o and GCP for the BharatBench OCR dataset. Shabdarth achieves the lowest 90 percentile for the Word Error Rate while being competitive with GCP on the Character Error Rate. Table 5 provides the performance of Chitrath against LLaMA 3.2-11B-V-Instruct on the translated academic datasets where Chitrath outperforms for different languages. We also provide baseline results on the BharatBench-V dataset in Table 6 and find Chitrath to be outperforming LLaMA 3.2-11B-V-Instruct for different languages.

4.4. Speech Understanding

4.4.1. Baselines

We consider the proprietary GCS as a baseline for Speech to text transcription task. In parallel, we also train custom models, based on Conformer (Gulati et al., 2020) like architecture on $\sim$ 2000 hours of data per language. We explore both Sarvam and Krutrim as the LLM backbone.

For the Speech-to-text translation tasks, we use Whisper model (Radford et al., 2023) as speech encoder, embeddings are mapped to text space using adapters. Here also, we explore both Sarvam and Krutrim as the LLM backbone. Models are trained on synthetically generated 16000 hours of speech to text translation dataset across 9 languages.

4.4.2. Results and Analysis

Based on the Word Error Rate (WER) metric, both the custom models perform better than GCS for different Indic languages on Speech-to-text transcription task. Krutrim as the LLM backbone performs better on Hindi, Gujarati and Kannada languages while Sarvam outperforms Krutrim for the speech to text translation on the BLEU metric.

5. Conclusion

Recent years saw an increased attention in building inclusive LLMs focusing on Indian languages. However, performance evaluation remains a key challenge. This work provides a comprehensive evaluation suite *BharatBench* across 8 Indian languages for different modalities including vision and speech. Through the release of our benchmarks and leaderboard, we intend to promote additional research and development in the evaluation of multilingual language models, enhancing cross-lingual NLP applications for Indian audience.

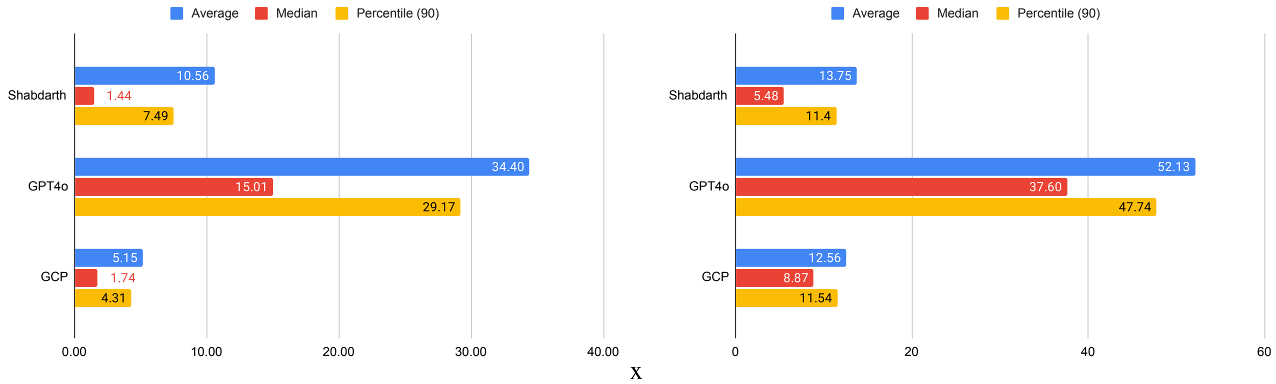
Limitations and Future Work: This is part of our ongoing effort to create a comprehensive evaluation suite. Our pilot study includes a collection of 8 Indic languages, which we plan to expand in the future iterations for the other low resource Indian languages. The selection of languages is in-

Bench	English	Telugu	Hindi	Bengali	Malayalam	Kannada	Tamil	Marathi	Gujarati	Assamese	Sanskrit
POPE											
Chitrarth	88	71	68	72	73	74	73	69	74	70	45
Llama3.2-11B-vision-instruct	88	55	75	63	54	65	61	64	66	72	34
LLaVA-Bench											
Chitrarth	68	55	52	54	56	58	58	53	56	59	59
Llama3.2-11B-vision-instruct	88	52	64	54	55	48	50	52	54	48	45
MM-Vet											
Chitrarth	39	44	39	33	25	46	34	41	39	37	41
Llama3.2-11B-vision-instruct	39	36	45	36	20	39	30	33	31	30	34

Table 5. **Performance of VLMs on BharatBench translated multi-modal academic datasets.**

Bench	Hindi	English	Telugu	Gujarati	Marathi	Tamil	Malayalam	Bengali	Kannada
Captioning									
Chitrarth	0.65	0.63	0.67	0.72	0.64	0.66	0.66	0.71	0.76
LLaMA 3.2V-11B-Instruct	0.65	0.62	0.63	0.66	0.59	0.58	0.72	0.65	0.64

Table 6. **Performance of baseline models on BharatBench-V Evaluation framework.**



(a) (b)

Figure 1. **We show Character Error Rate (CER) on the left and Word Error Rate (WER) on the right (lower is better).** Shabdarth model performs better against GPT-4o and GCP for the BharatBench OCR dataset

Model	Hindi	Bengali	Marathi	Gujarati	Tamil	Kannada	Telugu	Malayalam
GCS	19	29	21	22	26	26	35	33
Sarvam	14	17	15	19	22	19	27	27
Krutrim	13	18	16	18	22	18	27	29

Table 7. **Speech to text transcription measured through Word Error Rate (WER), lower is better.**

Model	Hindi	Bengali	Marathi	Gujarati	Tamil	Kannada	Telugu	Malayalam
Sarvam + LLama3.1	49	36	34	36	33	37	34	29
Krutrim	48	36	33	34	28	33	29	24

Table 8. **Speech to text translation measured using BLEU scores.**

formed by the availability of language-specific Indic models. Currently, the prompts employed for evaluation are limited in scope, and we intend to expand the number of prompts in subsequent iterations. For text-only LLM, we are working to expand our collection of tasks. We also want to include other modalities like videos as part of the future work. We consider our approach as a first step towards building general purpose multi-lingual evaluation framework which can handle various Indic languages and believe our research will foster improvements in multilingual LLM development and evaluation.

Authors List

Please cite this work as "BharatBench (2025)".

Language experiments: Guduru Manoj, Neel Rachamalla, Jay Piplodiya, Aditya Kallappa, Palash Kamble, Vivek Dahiya, Ashish Anand Kulkarni

Vision experiments: Shaharukh Khan, Ali Faraz, Akshat Patidar, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Shubham Agarwal, Abhinav Ravi

Speech experiments: Maitreyi M, Sanket Shah, Tejas Godambe, Nagaraj Adiga, Sharath Adavanne

Embedding experiments: Sandeep Pandey, Rajkiran Panuganti

Data, Evaluations, etc: Arveti Manjunath, Goutham Ramkumar, Bidyapathi Ray, Azhagiri S, Priyanka Nayak, Sandesh Bafna

Project Head: Chandra Khatri

Acknowledgements

We thank Bhavish Aggarwal and the rest of the Krutrim team which helped with the evaluation at various stages. Our models were trained with generous support from Krutrim cloud using Krutrim credits.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Adler, B., Agarwal, N., Aithal, A., Anh, D. H., Bhattacharya, P., Brundyn, A., Casper, J., Catanzaro, B., Clay, S., Cohen, J., et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

Aggarwal, D., Gupta, V., and Kunchukuttan, A. IndicXNLI: Evaluating multilingual inference for Indian languages. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10994–11006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.755. URL https://aclanthology.org/2022.emnlp-main.755.

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., and Sitaram, S. Mega: Multilingual evaluation of Generative AI. In *EMNLP 2023*, December 2023a.

Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Axmed, M., Bali, K., and Sitaram, S. Megaverse: Benchmarking large language models across languages, modalities, models and tasks, 2023b.

Artetxe, M., Ruder, S., and Yogatama, D. On the cross-lingual transferability of monolingual representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421.

Asai, A., Kasai, J., Clark, J., Lee, K., Choi, E., and Hajishirzi, H. XOR QA: Cross-lingual open-retrieval question answering. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 547–564, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.46. URL https://aclanthology.org/2021.naacl-main.46.

Balachandran, A. Tamil-llama: A new tamil language model based on llama 2, 2023.

Bendale, A., Sapienza, M., Ripplinger, S., Gibbs, S., Lee, J., and Mistry, P. Sutra: Scalable multilingual language model architecture. *arXiv preprint arXiv:2405.06694*, 2024.

Bhattacharjee, A., Hasan, T., Ahmad, W. U., Li, Y.-F., Kang, Y.-B., and Shahriyar, R. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2541–2564, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.143. URL https://aclanthology.org/2023.acl-long.143.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Choudhury, M., Chauhan, S., et al. Llama-3-nanda-10b-chat: An open generative large language model for hindi, 2024.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., and Kumar, P. 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)*, pp. 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL https://aclanthology.org/2023.acl-long.693.

Doddapaneni, S., Khan, M. S. U. R., Venkatesh, D., Dabre, R., Kunchukuttan, A., and Khapra, M. M. Cross-lingual auto evaluation for assessing multilingual llms. *arXiv preprint arXiv:2410.13394*, 2024a.

Doddapaneni, S., Khan, M. S. U. R., Verma, S., and Khapra, M. M. Finding blind spots in evaluator llms with interpretable checklists. *arXiv preprint arXiv:2406.13439*, 2024b.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dubois, Y., Liang, P., and Hashimoto, T. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.

Elo, A. E. and Sloan, S. The rating of chessplayers: Past and present, 1978.

FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., and Natarajan, P. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

Gala, J., Chitale, P. A., Raghavan, A. K., Gumma, V., Doddapaneni, S., M, A. K., Nawale, J. A., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., and Kunchukuttan, A. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=vfT4YuzAYA.

Gala, J., Jayakumar, T., Husain, J. A., M, A. K., Khan, M. S. U. R., Kanojia, D., Puduppully, R., Khapra, M. M., Dabre, R., Murthy, R., and Kunchukuttan, A. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv: 2401.15006*, 2024.

Gemma, T., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. Gemma: Open models based on gemini research and technology, 2024.

Google, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Gumma, V., Raghunath, A., Jain, M., and Sitaram, S. Health-pariksha: Assessing rag models for health chatbots in real-world multilingual settings. *arXiv preprint arXiv:2410.13671*, 2024.

Hada, R., Gumma, V., de Wynter, A., Diddee, H., Ahmed, M., Choudhury, M., Bali, K., and Sitaram, S. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*, 2023.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL `https://aclanthology.org/2020.findings-emnlp.445`.

Khan, M. S. U. R., Mehta, P., Sankar, A., Kumaravelan, U., Doddapaneni, S., Jain, S., Kunchukuttan, A., Kumar, P., Dabre, R., Khapra, M. M., et al. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv preprint arXiv:2403.06350*, 2024a.

Khan, S., Tarun, A., Ravi, A., Faraz, A., Pokala, P. K., Bhangare, A., Kolla, R., Khatri, C., and Agarwal, S. Chitrarth: Bridging vision and language for a billion people. In *NeurIPS Multimodal Algorithmic Reasoning workshop*, 2024b.

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., et al. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*, 2021.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Kohli, G. S., Parida, S., Sekhar, S., Saha, S., Nair, N. B., Agarwal, P., Khosla, S., Patiyal, K., and Dhal, D. Building a llama2-finetuned llm for odia language utilizing domain knowledge instruction set, 2023.

Krutrim, T. Krutrim LLM: Multilingual foundational model for over a billion people. *Under Review*, 2024.

Kumar, A., Shrotriya, H., Sahu, P., Mishra, A., Dabre, R., Puduppully, R., Kunchukuttan, A., Khapra, M. M., and Kumar, P. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5363–5394, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.360. URL `https://aclanthology.org/2022.emnlp-main.360`.

Kunchukuttan, A., Kakwani, D., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*, 2020.

Labs, T. Navarsa: Indic llms based on gemmma, 2023.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 5 2023a.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

Mhaske, A., Kedia, H., Doddapaneni, S., Khapra, M. M., Kumar, P., Murthy V, R., and Kunchukuttan, A. Naamapadam: a large-scale named entity annotated data for indic languages. *arXiv preprint arXiv:2212.10168*, 2022.

NLLB-Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation, 2022.

Pandey, S. and Panuganti, R. Vyakyarth: Towards leveraging small language models for indic embeddings. *Under submission*, 2025.

Parmar, J., Prabhumoye, S., Jennings, J., Patwary, M., Subramanian, S., Su, D., Zhu, C., Narayanan, D., Jhunjhunwala, A., Dattagupta, A., et al. Nemotron-4 15b technical report. *arXiv preprint arXiv:2402.16819*, 2024.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Rohera, P., Ginimav, C., Salunke, A., Sawant, G., and Joshi, R. L3cube-indicquest: A benchmark questing answering dataset for evaluating knowledge of llms in indic context. *arXiv preprint arXiv:2409.08706*, 2024.

Sarvam. Openhathi series: An approach to build bilingual llms frugally, December 2023. URL `https://www.sarvam.ai/blog/announcing-openhathi-series`.

Sarvam. Sarvam ai launches first llm for indian languages, October 2024. URL `https://www.sarvam.ai/blogs/sarvam-nvidia`.

Singh, A. K., Murthy, R., Sen, J., Ramakrishnan, G., et al. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. *arXiv preprint arXiv:2407.13522*, 2024a.

Singh, H., Gupta, N., Bharadwaj, S., Tewari, D., and Talukdar, P. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*, 2024b.

Sreenivas, S. T., Muralidharan, S., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., Catanzaro, B., Kautz, J., and Molchanov, P. Llm pruning and distillation in practice: The minitron approach. *arXiv preprint arXiv:2408.11796*, 2024.

Thellmann, K., Stadler, B., Fromm, M., Buschhoff, J. S., Jude, A., Barth, F., Leveling, J., Flores-Herr, N., Köhler, J., Jäkel, R., et al. Towards cross-lingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. MM-VET: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.