ParamBench: A Graduate-Level Benchmark for Evaluating LLM Understanding on Indic Subjects

Kaushal Sharma*♠, Vivek Patel*♠, Ayush Maheshwari and Aditya Maheshwari♠†

♣Indian Institute of Management Indore, India ♣BHARATGEN

{kaushals, vivekp, adityam}@iimidr.ac.in, ayush.hakmn@gmail.com

Abstract

Large language models (LLMs) have been widely evaluated on tasks such as comprehension, question answering, summarization, code generation, etc. However, their performance on graduate-level, culturally grounded questions in the Indian context remains largely unexplored. Existing Indian benchmarks emphasise basic factorientated queries that offer limited assessment of a deeper disciplinary understanding tailored to the Indian setting. In this paper, we present PARAMBENCH, consisting of around 11.5K questions in Hindi language comprising questionnaires from 16 diverse subjects. These questions are primarily derived from nation-wide graduate level entrance examination covering topics such as history, music, instruments, yoga, literature, philosophy, law, etc. ically for the Indian context. Additionally, we assess the ability of LLMs to handle diverse question formats—such as listbased matching, assertion-reason pairs, and sequence ordering—alongside conventional multiple-choice questions. We evaluated the performance of more than 17 open source LLMs on this benchmark, observing that Llama 3.3 70B attains the highest overall accuracy of 48%. Furthermore, subject-wise analysis indicates that even for the best performing LLMs, performance remains weak on topics such as music, classical instruments, politics and archaeology, underscoring persistent challenges in culturally grounded reasoning.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in multilingual

reasoning and knowledge-intensive tasks (Liu et al., 2024b). Although LLMs perform reasonably well in English and few other languages, their performance in culturally nuanced domains, particularly within the Indian context, remains weak (Verma et al., 2025). This is especially significant given India's linguistic and cultural diversity, with a population of over 1.4 billion, more than 120 major languages, and nearly 19,500 dialects across 28 states (Javed et al., 2024). Without robust evaluation in these settings, the application of LLMs to education, governance, and knowledge systems in India risks being incomplete and inequitable.

India has a rich body of traditional knowledge in areas such as history, religion, law, literature, philosophy, music, medicine, etc. Yet state-of-the-art LLMs often perform poorly when questions are related to familiarity with indigenous conceptual frameworks and knowledge (Maji et al., 2025). These weaknesses become clear in tasks that require an understanding of Indian ways of thinking, local concepts, and culturally specific knowledge. Existing Indic language benchmarks, while valuable for assessing syntactic and taskoriented competencies (Doddapaneni et al., 2023; Verma et al., 2025) fail to capture the diverse nuances. Recent resources such as the Sanskriti dataset (Maji et al., 2025) capture culturally salient attributes across India's geographic diversity. However, their emphasis on breadth leaves a gap for evaluation on indepth graduate level knowledge in culturally aligned subjects.

To address this gap, we present a new benchmark, Parambench consisting of roughly 11.5K questions across 16 India-focused subject areas—including archaeology, religion, law, culture, music, arts, philoso-

^{*} Authors contributing equally.

 $^{^{\}dagger}$ Corresponding author.

¹The dataset will be released soon. Please contact us if you'd like to evaluate openly available model.

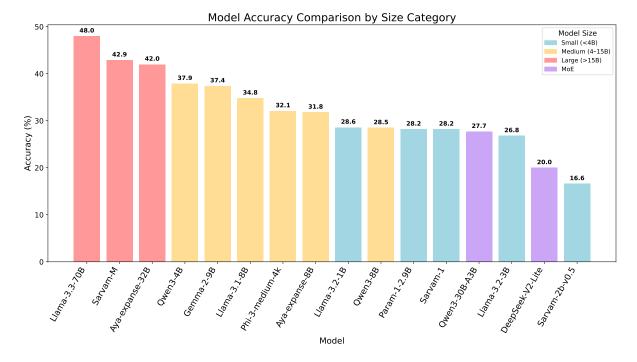


Figure 1: Average performance of evaluated models on PARAMBENCH, categorized by the parameter sizes.

phy, and yoga, etc. (c.f. Figure The 2). questions are drawn from postgraduate-level competitive exams and reflect fields grounded in India's intellectual and cultural traditions. Additionally, the benchmark includes multiple questions types: standard multiple-choice, list-based matching, assertion-reason, sequencing/ordering, incorrect-statement identification, and fill-in-the-blank. The benchmark tasks examines not only language understanding in Indian language but also whether models can understand and use concepts that are specific to Indian history, philosophy, law, literature, arts, etc.

In Parambench, we evaluate around 16 large language models of different model families and parameter sizes, including several recently released Indian LLMs. While LLMs often perform well on general-purpose tasks in English and on common topics, our results show a clear drop in performance on Indic topics (c.f. Figure 1). Among smaller models with fewer than 15 billion parameters, Qwen3-4B achieves the best overall accuracy at 37.9%. In the category of larger models, Llama-3.3-70B attains the highest accuracy of 48%. However, performance on several Indic subjects—including music, dance, philoso-

phy, archaeology, political science, and traditional instruments—remains below 41% even for the best-performing model. This highlights the persistent difficulty of culturally grounded subjects, where state-of-the-art open-source LLMs continue to struggle despite scale. Our contributions can be summarised as follows:

With Parambench, our aim is to identify and quantify current gaps in LLM performance for the Indian context and guide the development of models that are culturally and linguistically aligned with India. Our goal is to help build AI systems that better represent India's knowledge traditions and language diversity.

2 Related Works

2.1 LLM Benchmarks

Several benchmarks have been developed to evaluate the capabilities of large language models (LLMs). For instance, KMMLU and CMMLU (Son et al., 2025; Li et al., 2024) assess academic knowledge across a broad range of subjects, while BIG-Bench (Srivastava et al., 2023) focuses on measuring complex reasoning and generalization abilities. Similarly, HELM (Liang et al., 2023) introduces a comprehensive framework that evalu-

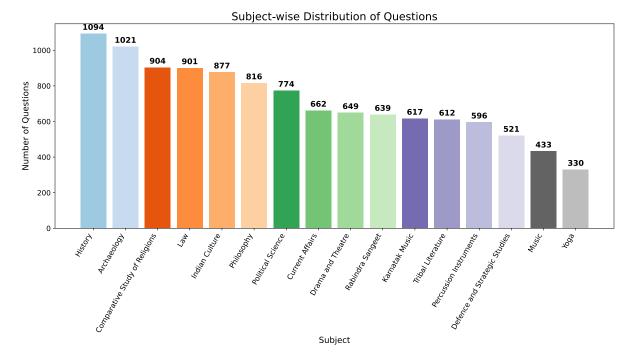


Figure 2: Distribution of the number of questions across different subjects

ates LLMs across multiple dimensions, including accuracy, robustness, and fairness.

Although these benchmarks provide extensive coverage of topics, they remain largely centered on English and other high-resource languages. Consequently, they often overlook linguistic and cultural diversity. Recent LLMs demonstrate some capacity for cultural and linguistic knowledge (Johnson et al., 2022; Atari et al., 2023; Masoud et al., 2025), yet continue to face significant challenges in adapting to non-Western contexts (Alkhamissi et al., 2024; Durmus et al., 2024).

2.2 Indian multilingual benchmarks:

In recent years, various benchmarks have been introduced to evaluate LLMs in the context of Indian languages and multilingual tasks. The Indic-QA Benchmark (Singh et al., 2025) provides large-scale question-answering datasets in 11 Indic languages, incorporating both original and translated content. MILU (Verma et al., 2025) expands this effort by presenting more than 80,000 multiple-choice questions in 11 languages, with a particular emphasis on culturally relevant topics. Similarly, IndicGenBench (Singh et al., 2024) focuses on generative tasks, such as summarization and translation, spanning 29 Indic lan-

guages. BharatBench (Krutrim, 2025) broadens the scope further by integrating text, vision, and speech modalities across 8 Indian languages.

More recently, cultural and evaluator alignment have been highlighted through benchmarks such as SANSKRITI (Maji et al., 2025) and PARIKSHA (Watts et al., 2024), which address the need to incorporate socio-cultural context in evaluating LLMs. Beyond task-specific benchmarks, several benchmarks have been proposed for extremely low-resource Indic languages such as Sanskrit (Maheshwari et al., 2022, 2024).

science and technical education, JEEBench (Arora et al., 2023) evaluates engineering entrance level mathematics, and the Materials Science Graduate Exam Benchmark targets post-graduate level scientific knowledge. In the legal and finance domain, IL-TUR (Joshi et al., 2024) assesses legal reasoning. while LLMs Acing Chartered Accountancy (Gupta et al., 2025) evaluates performance on taxation and auditing tasks. For governance and multilingual reasoning, datasets such as the UPSC Civil Services Study dataset (Banerjee et al., 2024), MILU (Verma et al., 2025), and IndicMMLU-Pro (KJ et al., 2025) examine general knowledge and reasoning across multiple Indic languages. Building on these efforts, our benchmark contributes by specifically evaluating India-centric knowledge through expert-verified multiple-choice questions (MCQs) drawn from the UGC-NET and UPSC examinations in Hindi.

3 ParamBench

In this section, we present data collection, annotation process and analysis.

3.1 Data Collection

Paramberch consists of 11,446 questions in Hindi language covering 16 Indic subjects such as Indian history, literature, archaeology, Indian culture, music, arts, yoga, etc. The subject-wise distribution of questions are present in Figure 2. The questions are collected from UGC-NET¹ and UPSC Civil services examination². UGC-NET is a nationwide examination administered by a government agency to determine eligibility for Ph.D. admission and for appointment to teaching positions in Indian universities and colleges. The exam is offered in around 80 subjects and conducted twice every year. Each test consists of two papers composed of multiple-choice questions (MCQs). UPSC Civil Services likewise employs rigorous multiple-choice assessments as part of a multi-stage selection process, providing domain-relevant, exam-realistic materials for evaluating graduate-level competence in India-specific subjects.

We constructed the dataset by downloading official question papers and answer keys from the respective examination websites. For UGC-NET, we curate papers from 2012–2018, selecting questions from 16 subjects that relates to Indian knowledge, including Indian history, law, music, and culture. We did not include other subjects as those are partially covered by other existing benchmarks such as Sanskriti (Maji et al., 2025), MILU (Verma et al., 2025). For UPSC, we include preliminary examination papers from 2011–2024, focusing on six major subjects that are central to Indian civilizational, literary, cultural, and academic knowledge. To the best of our knowl-

edge, this corpus has not appeared in prior LLM benchmarking studies and constitutes a newly curated, human-authored dataset designed explicitly for graduate-level evaluation in Indic contexts.

Each subject comprises of multiple question papers in PDF in which many of them are machine-readable, while a subset contains non-selectable text. Layouts vary across documents, with some in single-column format and the majority in two-column format. To ensure uniform text accessibility, we processed all PDFs with a proprietary OCR system and obtained text outputs for downstream curation and annotation.

3.2 Annotation setup

We process the OCR and extract text by tagging each question with its subject and the exam year of appearance. Following this, human annotators perform post-OCR correction to fix recognition errors and restore missing diacritics or script artifacts. Beyond textual corrections, annotators standardize formatting so that questions and answers are parseable by automated scripts. This pipeline ensures consistent metadata, clean text, and machine-actionable structure across heterogeneous source documents. Each entry was structured with fields for question, question type, options, correct answer, subject, year, and exam name to ensure consistency and traceability.

Annotation was conducted by subjectmatter experts proficient in Hindi and trained in the relevant domains. Because the source questions and answers are in Hindi, annotators were selected based on demonstrated fluency in reading, writing and speaking of Hindi and grammar knowledge.

General annotation guidelines were developed and shared with all annotators. These emphasized grammatical correctness, completeness of questions and answers, and standard formatting. Questions with unresolved issues were to be removed or escalated for review. Annotation was primarily carried out using Google Docs, and the finalized dataset was exported in CSV format.

¹https://ugcnet.nta.ac.in

²https://upsc.gov.in/examinations

3.3 Team structure

The annotation team consisted of two tiers. First, a team of four annotators corrected questions and answers from OCR outputs, entered answer keys, and corrected grammatical errors. This was followed by a review process by subject expert, who verified the grammar, formatting, and correctness of the answer keys.

We implemented both manual and automated quality assurance protocols. Initially, only those questions that aligned with the benchmark's focus on Indian-specific knowledge were retained. Manual checks were performed to validate grammatical correctness, answer accuracy, and completeness. Automated scripts were used to ensure each question had exactly four options, one correct answer, and no missing fields. This two-step quality control process helped maintain the reliability of the annotated dataset.

Prior to full-scale annotation, annotators were given a sample dataset along with worked examples. They were then assigned trial files, which were reviewed and corrected by the reviewers. Feedback was provided iteratively, and only after demonstrating consistent accuracy were annotators assigned larger batches of data. All annotation work was performed using Google Spreadsheets. Annotators were compensated at a rate of \$1 per 10 questions.

Question Type	# Questions
Normal MCQ	7393
Find incorrect Statement	1250
Match the List	1129
Assertion and Reason	1001
Sequencing	646
Fill in the Blank	27
Total	11446

Table 1: Distribution of question types in Param-Bench

3.4 Statistics

Figure 2 presents the overall distribution of questions in Parambench. The corpus comprises 11,446 Hindi questions spanning 16 subjects. History contributes the largest share (1,094 questions), whereas Yoga has the fewest (330 questions). Notably, the benchmark includes substantial coverage of Indic domains

that are rarely represented in prior evaluations, such as drama and theater, Rabindra sangeet, Tribal and Regional Literature, Percussion Instruments, and Yoga. Although some categories exhibit thematic overlap—for example, Music with Carnatic music and Rabindra sangeet—we retain them as distinct subjects to preserve domain specificity and enable fine-grained analysis.

We categorize each item in the table by question type to reflect the range of exam-style reasoning (see Table 1). The benchmark is primarily composed of standard multiple-choice questions, with additional formats such as list-based matching, assertion—reason pairs, sequencing/ordering, incorrect-statement identification, and a small set of fill-in-the-blank items. Overall, MCQs form the majority, while the other types provide complementary coverage of mapping, causal reasoning, temporal ordering, and verification skills.

4 Experiments

We evaluate Parambench on 16 openly available models spanning diverse sizes and architectures, including parameter ranges of 1B–4B, 4B–15B, and over 15B, as well as two Mixture-of-Experts (MoE) models (Table 2). All models are instruction-tuned variants (see Section 4.1). All experiments follow a zero-shot setup using the prompt in Table 8, without demonstrations or validation examples. Models are executed locally through using Hugging Face transformers library.

Evaluation is based on direct answer matching generated responses are compared with the gold key, and accuracy is reported as the proportion of correct outputs. This yields a clear and interpretable measure of task performance in realistic settings.

4.1 Evaluation setup

We evaluated a range of open-source large language models (LLMs), spanning small-scale (1B parameters) to 70B parameters models. The selection includes both base and instruction-tuned variants across different architectures and training. The following models were used in our evaluation:

1. **Meta Llama Series:** We have used the Llama 3 collection, which offers several mul-

tilingual language models (Team, 2024b). We used Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, Llama3.3-70B during experiments.

- 2. Sarvam Models: Sarvam-1 and Sarvam-2b-v0.5 are 2B, 2.5B parameter language models respectively. We also evaluate Sarvam-M which is a 24B parameter model post-trained on Mistral-Small-3.1(SarvamAI).
- 3. **Param-1-2.9B:** PARAM-1 a bilingual language model trained from scratch in English and Hindi containing 2.9 billion parameters (Pundalik et al., 2025).
- 4. **Qwen3 series:** This includes model Qwen3-4B, Qwen3-8B dense model. We also evaluate on MoE Qwen3-30B-A3B containing 30B total parameters and 3B active parameters (Team-Qwen3, 2025).
- 5. **Gemma-2-9B:** This model is a 9B parameter instruction-tuned language model (Team-Gemma2, 2024).
- 6. **Phi-3-medium-4K:** This model is a 14B parameter open multilingual model (Team, 2024c).
- 7. Cohere series: Aya 23-8B and Aya-23-35B are instruction fine-tuned model with multilingual capabilities. (Aryabumi et al., 2024).
- 8. **DeepSeek-V2-Lite:** This model is a lighter version of an original DeepSeek model containing 16B total parameters and 2.4B active parameters (Liu et al., 2024a).

5 Result

5.1 Model results

The evaluation results in the Table 2 shows that overall accuracies of LLMs models remain comparate vely low across all parameter sizes. In the category of models with fewer than 4B parameters, the highest performance was observed with Llama-3.2-1B (28.6%) (Team, 2024b), while most others, including Param-1-2.9B (Pundalik et al., 2025) and Sarvam-1, stayed around 28 %. Among mid-sized models (4B–15B parameters), performance improved significantly, with Qwen3-4B (Team-Qwen3, 2025) (37.9 %) and Gemma-2-9B (Team-Gemma2, 2024) (37.4 %) performing best. Larger models with more than 15B parameters achieved comparatively higher scores, led by Llama-3.3-70B (Team, 2024b) at 48.0 %, but even this top result falls short of a majoritylevel accuracy.

Model	Acc (%)			
< 4B params				
Llama-3.2-1B	28.6			
Sarvam-1	28.2			
Sarvam-2b-v0.5	16.6			
Param-1-2.9B	28.2			
Llama-3.2-3B	26.8			
4B-15B pa	rams			
Qwen3-4B	37.9			
Qwen3-8B	28.5			
Llama-3.1-8B	34.8			
Aya-expanse-8B	31.8			
Gemma-2-9B	37.4			
Phi-3-medium-4k	32.1			
> 15B par	rams			
Sarvam-M	42.9			
Aya-expanse-32B	42			
Llama-3.3-70B	48			
MOE				
DeepSeek-V2-Lite	20			
Qwen3-30B-A3B	27.7			

Table 2: Paramberch evaluation results of openly available models for different parameter sizes. The numbers are averaged across all subjects and reported as accuracy in percentage.

5.2 Subject wise results

Table 3 present the zero-shot evaluation performance of models with lower than 4B parameters. Sarvam-1 (SarvamAI) shows good results in Comparative Study of Religions (41.6 %) and Political Science (31.8 %), while Llama-3B (Team, 2024b) achieves the best accuracy in Defence and Strategic Studies (35.3 %) and Tribal Literature (34.3%). Llama-1B (Team, 2024b) also performs well in Yoga (33.6 %) and Defence (33 %). Param-1 (Pundalik et al., 2025) records competitive scores in subjects such as Comparative Study of Religions (32.9 %) and Tribal Literature (32.4 %). Alternatively, Sarvam-2B usually lags behind the others, though it performs relatively better in Yoga (25.5 %). Concluding, Sarvam-1 and Llama-3B (Team, 2024b) appear as stronger performers across various subjects, whereas Sarvam-2B shows weaker consistency.

For models in the 4B–15B parameter range, performance is stronger and more balanced across subjects compared to the smaller models, as shown in the Table 4. Gemma-2

${\bf Models\ with < 4B\ parameters}$									
	Subject Param-1 Llama-1B Llama-3B Sarvam-1 Sarvam-2								
Archaeology	29.3	30.5	23.4	25.4	16.4				
Comparative Study of Religions	32.9	32.1	35.3	41.6	22.4				
Current Affairs	$\frac{32.9}{27.8}$	28.7	28.7	29.2					
0 00 000					16				
Defence and Strategic Studies	32.6	33	35.3	34	15.6				
Drama and Theatre	25.1	29.6	21.3	25.3	19.9				
History	27.4	28.1	24	25.6	14.1				
Indian Culture	25.1	24.7	25.4	26.2	14.9				
Karnatak Music	28.9	26.6	27.2	24.2	16.9				
Law	27.6	29	26.1	27.3	15.3				
Music	26.3	23.1	24	24.9	18.9				
Percussion Instruments	25.8	29.2	26.9	24.7	14.6				
Philosophy	23.7	27	23.2	28.4	16.9				
Political Science	27.4	26	23.3	31.8	12.8				
Rabindra Sangeet	31.6	26.5	26.6	23.9	11				
Tribal Literature	32.4	32.4	34.3	29.7	22.1				
Yoga	29.4	33.6	28.8	24.6	25.5				

Table 3: Subject-wise results of models with $<\!4\mathrm{B}$ parameters. Numbers are reported as accuracy in percentage.

Models with 4B–15B parameters						
Subject	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Aya-8b	Gemma-2	Phi-3
Archaeology	34.4	25	33	28.2	30.9	31
Comparative Study of Religions	48.6	32.2	48.9	45.9	54.9	34.9
Current Affairs	47.6	36.6	40.9	35.5	52	37.8
Defence and Strategic Studies	47.8	30.5	42.2	40.9	49.1	36.5
Drama and theatre	35.6	29	31.7	32.7	29	33
History	35.1	26.3	31.8	29.1	36.2	28.9
Indian Culture	33.2	27.3	31.8	25.1	33.9	30.6
Karnatak Music	32.6	28.4	28.5	28.4	28.7	29.2
Law	35.3	24.5	33.1	29	35.6	28
Music	30.7	27.5	24.7	26.3	24.3	30.3
Percussion Instruments	30.4	29.4	32.1	28.9	34.1	31.4
Philosophy	40.4	25.9	35.4	33.3	37.8	31.7
Political Science	36.8	25.6	32.4	28.6	34.6	31.3
Rabindra Sangeet	33.7	26.3	29.7	25	29.7	34.4
Tribal Literature	46.4	37.8	44.4	40.5	49.2	38.7
Yoga	39.7	29.4	31.8	35.5	35.2	28.8

Table 4: Subject-wise results of models with 4B-15B parameters. Numbers are reported as accuracy in percentage.

m Models~with > 15B~parameters~and~MoE						
Subject	Sarvam-M	Aya-32B	Llama-3.3-70B	DeepSeek	Qwen3-30B-A3B	
Archaeology	39.5	33.5	40.9	17.1	28.2	
Comparative Study of Religions	57.9	62.9	67.1	23.6	32.2	
Current Affairs	63.9	55	69	21.5	38.1	
Defence and Strategic Studies	54.9	56.8	66.8	23	32.1	
Drama and theatre	36.2	37	42.2	24.2	25.3	
History	43	38.5	46.5	19	27.4	
Indian Culture	40.6	39.9	49.3	18.8	25.8	
Karnatak Music	32.6	32.3	33.6	21.9	28	
Law	37.2	36.6	43	21.8	26.5	
Music	33.3	35.6	35.1	18.2	23.8	
Percussion Instruments	29.5	32.2	35.1	22	28.0	
Philosophy	43.8	43.8	49.6	15.4	19.2	
Political Science	42.8	43.8	47.4	17.1	27.3	
Rabindra Sangeet	32.2	33	32.9	19.9	27.1	
Tribal Literature	52.1	52.1	58.7	22.6	30.9	
Yoga	45.8	37	48.2	13.9	25.8	

Table 5: Subject-wise results of models with >15B parameters. Numbers are reported as accuracy in percentage.

persistently stands out, achieving the highest scores in Comparative Study of Religions (54.9 %), Current Affairs (52 %), and Defence and Strategic Studies (49.1 %). Qwen3-4B (Team-Qwen3, 2025) also functions well, specifically in Comparative Study of Religions (48.6 %) and Defence (47.8 %). Llama-3.1-8B (Team, 2024b) shows competitive results in subjects like Comparative Study of Religions (48.9 %) and Tribal Literature (44.4 %), while Aya-8B (Aryabumi et al., 2024) records strong performance in Comparative Study of Religions (45.9 %) and contributes steady scores across multiple subjects. Phi-3 (Aryabumi et al., 2024) usually stays in the mid-range but achieves its best in Tribal Literature (38.7) and Rabindra Sangeet (34.4). Overall, Gemma-2 emerges as the strongest performer, with Qwen3-4B (Team-Qwen3, 2025) and Llama-3.1-8B (Team, 2024b) also showing robust results, particularly in knowledgeheavy subjects.

For models with more than 15B parameters and MoE architectures, performance reaches its strongest levels overall as shown in Table 5. Llama-3.3-70B (Team, 2024b), by obtaining highest scores in Current Affairs (69 %), Comparative Study of Religions (67.1 %), Defence and Strategic Studies (66.8), and Tribal Literature (58.7 %), consistently dominates all subjects. Aya-32B (Aryabumi et al., 2024) also performs competitively, in Comparative Study of Religions (62.9 %) and Defence (56.8 %), while Sarvam-M (SarvamAI) presents strong

results in Current Affairs (63.9 %) and Comparative Study of Religions (57.9 %). Alternatively, the MoE models (DeepSeek (Team, 2024a) and Qwen3-30B-A3B (Team-Qwen3, 2025)) fall behind, with DeepSeek putting forward weak performance across most domains while Qwen3-30B achieving only slight improvements, by performing its best in Current Affairs (38.1 %). Overall, Llama-3.3-70B clearly emerges as the best performer, with Aya-32B and Sarvam-M showcasing strong choices, but MoE models underperformed in this evaluation.

Models with fewer than 4B parameters show only scattered and limited results with Sarvam-1 (SarvamAI) and Llama-3B (Team, 2024b) providing some advantages for a few subjects. While performance improves significantly, with Gemma-2 (Team-Gemma2, 2024), Qwen3-4B (Team-Qwen3, 2025), and Llama-3.1-8B (Team, 2024b) achieving strong and balanced performance in knowledge-intensive subjects such as Comparative Study of Religions, Current Affairs, and Defence Stud-Most interestingly, performance visible for models greater than 15B parameters, whereby Llama-3.3-70B achieves the best performance across almost all subjects, with Aya-32B (Aryabumi et al., 2024) and Sarvam-M (SarvamAI) being dictated by somewhat lower Also, the Mixture-of-Experts performance. (MoE) models appear to underperform relative to their similar-scale dense model counterparts, revealing current shortcomings of MoE models in this benchmark. Overall, the results imply a strong relationship with the size of parameters to performance, whereby large dense models exhibit the most reliable accuracy across unitary subject areas.

5.3 Question type wise results

The performance varied widely among the models across the question types, with larger models achieving in general a higher accuracy as present in Table 6. Llama-3.3-70B (Team, 2024b) recorded the best results overall, which exceeded 50% in both normal MCQ and assertion-reason questions, and performed strongly across most of the cate-Aya-expanse-32B (Aryabumi et al., 2024) and Sarvam-M also achieved high scores, specifically in normal MCQ and identifythe-incorrect-statement tasks. The mid-sized models like Qwen3-4B (Team-Qwen3, 2025) and Gemma-2-9B (Team-Gemma2, 2024) delivered competitive results, especially in formats which included normal MCQ and assertion-reason type questions, while MOE model such as DeepSeek-V2-Lite (Liu et al., 2024a) had significantly low accuracies, which were often below 20% in several categories. Tasks like match-the-list and fill-in-the-blank showed a tendency to have lower accuracy overall, which indicated that they posed challenges across most model sizes.

6 Conclusion

LLMs continue to fall short when evaluated on culturally grounded, India-specific domains despite strong performance on general benchmarks. Parambench fills this gap by offering a rigorous, graduate-level evaluation across diverse subjects rooted in India's intellectual traditions. Our results reveal clear performance drops across leading models, emphasizing the urgent need for culturally aligned benchmarks. We envision Parambench as both a diagnostic tool and a stepping stone toward developing LLMs that are more inclusive of India's linguistic and knowledge diversity.

Acknowledgments

We thank the Bharatgen initiative for fostering a collective effort toward developing foundation models that reflect India's linguistic and cultural diversity. This work aligns with Bharatgen's broader mission of building inclusive and representative AI ecosystems, and we are grateful to be part of this shared endeavour.

References

Badr Alkhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12404–12422.

Daman Arora, Himanshu Singh, et al. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7527–7543.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and So*cial Psychology, 125(5):1157–1188.

Somonnoy Banerjee, Sujan Dutta, Soumyajit Datta, and Ashiqur R KhudaBukhsh. 2024. Gender representation and bias in indian civil service mock interviews. *CoRR*.

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.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2024. Towards measuring the representation of subjective global opinions in language models. In First Conference on Language Modeling.

Jatin Gupta, Akhil Sharma, Saransh Singhania, Mohammad Adnan, Sakshi Deo, Ali Imam

Models	MCQ	IS	$\mathbf{L}\mathbf{M}$	A&R	Sequence	\mathbf{BF}
Llama-3.2-1B	28.3	21.9	29.8	39.8	26.5	14.8
Sarvam-1	30.4	28	16.7	28.5	22.1	33.3
Sarvam-2b-v0.5	16.1	15.9	18.7	18.9	17.4	29.6
Param-1-2.9B	29.1	20.5	28	32.5	27.3	22.2
Llama-3.2-3B	28.7	21.9	14.1	33.5	25.2	18.5
Qwen3-4B	37.8	40.2	34	41.9	35.5	40.7
Qwen3-8B	30	27.2	7.3	38.6	31.3	25.9
Llama-3.1-8B	37.7	27.2	28.1	34.3	28.2	37
Aya-expanse-8B	34.2	25	21.6	37.9	25.7	25.9
Gemma-2-9B	39.3	33.3	32.3	40.3	29.5	25.9
Phi-3-medium-4k	33.6	33	25.8	27.8	29.7	33.3
Sarvam-M	45.9	39.7	36.6	37.5	35.5	37
Aya-expanse-32B	46	42.2	20.4	44.7	28.5	25.9
Llama-3.3-70B	52	38.6	32.7	53	39.2	40.7
${\bf Deep Seek\text{-}V2\text{-}Lite}$	20.3	17.9	22.9	20	15.8	14.8
Qwen3-30B-A3B	26.4	24.8	28.7	40.9	27	25.9

Table 6: Average accuracy across all subjects for different question types. In this table, IS represents Find Incorrect Statement, LM represents Match the List, A&R represents Assertion and Reason, BF refers to Fill in the blanks, sequence refers to Correct sequence ordering and MCQ refers to Multiple choice questions.

Abidi, and Keshav Gupta. 2025. Large language models acing chartered accountancy.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, et al. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. In Findings of the Association for Computational Linguistics ACL 2024, pages 10740–10782.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. Il-tur: Benchmark for indian legal text understanding and reasoning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11460–11499.

Sankalp KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding. *CoRR*, abs/2501.15747.

Team Krutrim. 2025. Bharatbench: Comprehensive multilingual multimodal evaluations of foundation ai models for indian languages.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Maosong Sun, and Erhong Yang. 2024b. Omgeval: An open multilingual generative evaluation benchmark for large language models. arXiv preprint arXiv:2402.13524.

Ayush Maheshwari, Ashim Gupta, Amrith Krishna, Atul Kumar Singh, Ganesh Ramakrishnan, Anil Kumar Gourishetty, and Jitin Singla. 2024. Samayik: A benchmark and dataset for English-Sanskrit translation.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-OCR text correction in Sanskrit.

- Arijit Maji, Raghvendra Kumar, Akash Ghosh, Sriparna Saha, et al. 2025. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. arXiv preprint arXiv:2506.15355.
- RI Masoud, Z Liu, M Ferianc, P Treleaven, and M Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Proceedings-International Conference on Computational Linguistics, COLING, pages 8474–8503. Association for Computational Linguistics (ACL).
- Kundeshwar Pundalik, Piyush Sawarkar, Nihar Sahoo, Abhishek Shinde, Prateek Chanda, Vedant Goswami, Ajay Nagpal, Atul Singh, Viraj Thakur, Vijay Dewane, Aamod Thakur, Bhargav Patel, Smita Gautam, Bhagwan Panditi, Shyam Pawar, Madhav Kotcha, Suraj Racha, Saral Sureka, Pankaj Singh, Rishi Bal, Rohit Saluja, and Ganesh Ramakrishnan. 2025. Param-1 bharatgen 2.9b model.
- Sarvam AI | Sovereign Indian AI | Ecosystem for LLMs, Agents, and AI Assistants sarvam.ai. https://www.sarvam.ai/. [Accessed 20-08-2025].
- 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. 2024. 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.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. Kmmlu: Measuring massive multitask language understanding in korean. 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 4076–4104.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: quantifying and extrapolating the capabilities of language

- models. Transactions on Machine Learning Research, 2023(5):1–95.
- DeepSeek-AI Team. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.
- Llama 3 Team. 2024b. The llama 3 herd of models.
- Phi-3 Team. 2024c. Phi-3 technical report: A highly capable language model locally on your phone.
- Team-Gemma 2: Improving open language models at a practical size.
- Team-Qwen3. 2025. Qwen3 technical report.
- Sshubam Verma, Mohammed Safi Ur Rahman, Vishwajeet Kumar, Rudra Murthy Venkataramana, and Jaydeep Sen. 2025. Milu: A multitask indic language understanding benchmark. In Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7900–7932.

7 Appendix

Examples: Types of questions in ParamBench

The following table 7 presents examples of six distinct types of questions used in our benchmark dataset. For each type, two representative questions have been selected to illustrate the structure, content, and answer format.

Type	Question	Options	Ans.
Normal MCQ	पाषणकालीन उपकरणों की उपयोगिता की अध्ययन विधि है :	(a) स्तर विज्ञान (b) सूक्ष्म चिह्नीय अध्ययन (c) शुल्कन प्रयोग (d) प्रारूपकीय विज्ञान	(b)
Normal MCQ	निम्नांकित मृदभांड परम्पराओं में से कौन महाभारत काल से जुड़ा हुआ है ?	(a) उत्तरी काले चमकीले मृदभांड (b) कृष्ण-लोहित मृदभांड (c) गैरिक मृदभांड (d) चित्रित धूसर मृदभांड	(d)
Incorrect Statement Identification	उस कूट को चिन्हित करें जिसमें सही अभिकथन न हो :	(a) एक संयोजक सत्य है यदि इसके सभी संघटक सत्य हैं अन्यथा यह असत्य है। (b) प्रत्येक यौगिक अभिकथन एक सत्यता– फलन अभिकथन होता है। (c) द्विमूल्याश्रित तर्कशास्त्र में प्रत्येक अभिकथन या तो सत्य होता हैं या असत्य। (d) एक सरल अभिकथन वह अभिकथन है जिसका संघटक इसके भाग के रूप में अन्य कोई अभिकथन नहीं होता।	(b)
Incorrect Statement Identification	निम्नलिखित में से कौन सा युग्म सही सुमेलित नहीं है ?	(a) खारवेल का हाथीगुम्फा — पार्श्वनाथ अभिलेख (b) चन्द्रगुप्त द्वितीय का — वीरसेन उदयगिरि गुफा अभिलेख साब (c) रुद्रदामन का — तुषास्फ जूनागढ़ शिलालेख (d) कुमारगुप्त एवं — वत्सभिट्ट बन्धुवर्मा का मन्दसौर प्रस्तर अभिलेख	(a)
List-based Matching	सूची-। को सूची-॥ के साथ सुमेलित करें। सूची-।: (a) पक्षधर्मता, (b) विपक्षसत्त्व, (c) बाधित, (d) विरुद्ध सूची-॥: (i) अग्नि शीतल है, (ii) शब्द शाश्वत है क्योंकि यह उत्पन्न होता है, (iii) पर्वत पर धूम्र है, (iv) जलाशय में अग्नि है	(a) (iv) (i) (iii) (ii) (b) (i) (ii) (iii) (iv) (c) (ii) (iii) (i) (iv) (d) (iii) (iv) (i) (ii)	(d)

Type	Question	Options	Ans.
List-based	सूची-। और सूची-॥ को सुमेलित करें।	(a) (iv) (iii) (ii) (i)	(c)
Matching	सूची-I: (a) वैयक्तिक प्रत्ययवाद, (b) एकतत्त्व	(b) (ii) (iv) (i) (iii)	
	प्रत्ययवाद, (c) आत्मनिष्ठ प्रत्ययवाद, (d)	(c) (ii) (i) (iv) (iii)	
	यथार्थवादी प्रत्ययवाद	(d) (i) (ii) (iii) (iv)	
	सूची-II: (i) सीमित आत्मा एक का अंश,		
	प्रकार अथवा अभास है। (ii) मूर्तसत्ता वैयक्तिक		
	आत्मत्व है। (iii) वस्तुओं के आदर्श रहित रूपों		
	की यथार्थता को स्थापित करते हैं। (iv) प्रकृति		
	सीमित मन का प्रक्षेपण मात्र है।		
Assertion &	दिये गये अभिकथन (A) और तर्क (R) की	(a) दोनों (A) और (R) सही	(c)
Reasoning	परीक्षा आगमनात्मक अनुमान के आलोक में करें	व्याख्या है	
	और नीचे दिये गये कूट में से सही का चयन	(b) दोनों (A) और (R) सही	
	करें। अभिकथन (A) : आगमनात्मक अनुमान में	व्याख्या नहीं है	
	ज्ञात से अज्ञात की ओर जाते हैं। तर्क (R) :	(c) (A) सही, (R) गलत	
	आगमनात्मक अनुमान में किसी जाति विशेष के	(d) (R) सही, (A) गलत	
	सभी सदस्यों के निर्णय के द्वारा उस जाति विशेष		
	के सभी सदस्यों के बारे में निर्णय तक पहुँचते हैं।		
	कूट :		
Assertion &	नींचे एक अभिकथन (A) और एक कारण	(a) (A) और (R) दोनों सही	(a)
Reasoning	(R) दिये गये हैं। उन पर विचार कीजिये और	हैं और (R), (A) का सही	
_	नीचे दिये गये कूट से सही विकल्प का चयन	आधार है।	
	कीजिये। अभिकथन (A): परमाणु की सत्ता	(b) (A) और (R) दोनों सही	
	अवश्य स्वीकार की जानी चाहिये। तर्क (R):	हैं और (R), (A) का सही	
	द्वयणुक सावयव है। कूट :	आधार नहीं है।	
	,	(c) (A) सही है और (R)	
		गलत है।	
		(d) (A) गलत है और (R)	
	\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\	सही	
Sequencing	कालसमयानुसार ग्रन्थों का सही क्रम चुनिए :	(a) संगीत मकरंद, राग-	(d)
		विबोध, नारदीय–शिक्षा,	
		राग-तरंगिनी	
		(b) नारदीय-शिक्षा, संगीत	
		मकरंद, राग–विबोध, राग– तरंगिनी	
		तरागना (c) राग–तरंगिनी, संगीत	
		(८) राग-तरागना, संगात मकरंद, नारदीय–शिक्षा,	
		नकरद, नारदाय=।राक्षा, राग=विबोध	
		(d) नारदीय–शिक्षा, संगीत	
		(व) गारपाय गरापा, संगारा मकरंद, राग–तरंगिनी,	
		राग=विबोध	
		11-14414	

Type	Question	Options	Ans.
Sequencing	कालक्रमानुसार सही क्रम चुनिए :	(a) खुदा बक्श, फैयाज़ खान,	(b)
		गुलाम अब्बास, शराफत हुसैन	
		(b) खुदा बक्श, गुलाम	
		अब्बास, फैयाज़ खान,	
		शराफत हुसैन	
		(c) फैयाज़ खान, शराफत	
		हुसैन, खुदा बक्श, गुलाम	
		अब्बास	
		(d) शराफत हुसैन, खुदा	
		बक्श, फैयाज़ खान, गुलाम	
		अब्बास	
Fill in the	सखी-कुंधे नट की रंगमंचीय प्रस्तुति	(a) असम	(d)
blanks	है।	(b) आंध्र प्रदेश	
		(c) पंजाब	
		(d) ओडिशा	
Fill in the	निम्नलिखित में रिक्त स्थान की पूर्ति करें:	(a) के समकक्ष	(c)
blanks	अन्तःप्रज्ञावाद, अन्तःप्रज्ञात्मता	(b) के समान	
		(c) से भिन्न	
		(d) इनमें से कोई नहीं	

Prompt = f"""Question: {['question_text']} Options: A) {['option_a']} B) {['option_b']} C) {['option_c']} D) {['option_d']} Given the above question and multiple options, select the correct answer. Keep your response only in English with one of the letters corresponding to the options A, B, C, or D. Do not write

Table 8: Zero-Shot prompt applied across all models for evaluation

anything else.""".