New proposed reasoning benchmarks for Indic languages

Addressing contamination, toxicity, and missing reasoning capabilities

Selim Khouaja

Infosys, Imperial College London

July 21, 2025

Outline

Current SOTA of Benchmarking in Indic Languages

Benchmark	Langs	Task Types	Best Perf.	Coverage Area
IndicMMLU-Pro	9	Multiple-choice QA across 57 subjects translated from MMLU-Pro	44.80% (GPT-4o)	Knowledge recall across domains
IndicGenBench	29	Cross-lingual summarization, machine translation, multilingual QA	Variable	Text generation tasks
MILU	11	Multiple-choice QA across 8 domains (STEM, Arts, Law, Business) from national/state exams	74% (GPT-4o)	India-specific knowledge
UNITYAI-GUARD	6	Binary toxicity classification using 14-category Llama-Guard taxonomy	84.23% F1	Content moderation

Current SOTA of Benchmarking in Indic Languages - 2

Benchmark	Langs	Task Types	Best Perf.	Coverage Area
PARIKSHA	10	Pairwise comparison and direct assessment using culturally-nuanced prompts	90K evals	Human-LLM alignment
MEGAVERSE	14	Translation, QA, summarization, sentiment analysis, NER, classification	GPT-4 best	Cross-lingual evaluation
IndicXTREME	22	Natural language inference, QA, classification, structured prediction	76% (MuRIL)	Language understanding

Performance Analysis Summary

- **Understanding Tasks**: 44-76% accuracy range, significant gap from human baselines
- **Generation Tasks**: Consistent performance degradation vs English across all 29 languages in IndicGenBench
- **Cultural Knowledge**: MILU's 74% peak shows average grasp of India-specific contexts
- **Toxicity Detection**: UNITYAI-GUARD achieves strongest performance at 84.23% F1, but the benchmark has limited scope
- **Human Evaluation**: PARIKSHA shows human-LLM agreement varies significantly across languages

Performance Analysis Summary - 2

Critical Missing Capabilities

- **Open-ended Reasoning**: No Indic benchmark treats the abilities of LLMs to answer long dissertation style questions
- **Language Skills Generalization Abilities**: No Indic benchmark treats the broad generalization abilities of LLMs in language tasks. We believe this can be achived by measuring their abilities in puzzles.
- **Temporal-Spatial Integration**: No benchmark treats the abilities of LLMs to combine temporal and spatial constraints to serve as planners or navigator agent
- **Analytical Reasoning**: No Indic benchmark treats the abilities of LLMs to interpret data and draw conclusions.
- **Creative Problem-Solving**: No Indic benchmark treats the abilities of LLMs to construct stories produce art, etc.

Key Limitations - Contamination and Data Quality

Benchmark	Contamination Risk
IndicMMLU-Pro	Machine-translated from English MMLU-Pro, likely in training data; measures pattern memorization rather than genuine understanding
IndicGenBench	Extends CrossSum, FLORES, XQuAD datasets widely used in training; inflates cross-lingual generation performance
MILU	Web-scraped from public exam portals and educational websites; compromises India-specific knowledge assessment
UNITYAI-GUARD	Only 30k manually verified test vs 567k training instances; limited contamination detection
PARIKSHA	Human-curated prompts created specifically for evaluation; maintains reliability
MEGAVERSE	Aggregates multiple translated English benchmarks; cross-lingual contamination
IndicXTREME	Extensions of XTREME benchmark components; English source leakage

Key Limitations - Missing Reasoning and Evaluation Capabilities

Missing Reasoning Type	What This Means for AI Assessment
Open-ended Reasoning	Cannot assess creative problem-solving, causal chain analysis, or analogical thinking where models must generate novel solutions rather than selecting from predefined options
Language Puzzles/Generalization	No evaluation of compositional understanding through crosswords, anagrams, riddles, or word games that test ability to combine linguistic elements in novel ways
Temporal-Spatial Integration	Missing assessment of reasoning that combines time-based sequencing with geographical/spatial constraints, like planning multi-step journeys or coordinating events across locations
Analytical Reasoning	Lack of quantitative analysis tasks requiring data interpretation, statistical inference, logical argument construction, and evidence synthesis from multiple sources
Creative Reasoning	No measurement of innovation, artistic generation, or creative problem-solving that requires models to produce original content rather than reproduce learned patterns

Requirements - Contamination-Free Design

To ensure our benchmark does not create data leakage and suffer from contamination, our dataset of tasks must be

Contamination Prevention

- **Date Controlled**: Control for all contents/questions to be passed a certain date D. This is still to be determined such as to provide enough content but be as late as possible
- **Source Controlled**: Need to establish a trusted and accessible list of sources to extract the datasets content from
- **Validated**: Need to implement a contamination detection test, preventing contamination from past used datasets or older datasets
- **Protected**: Need to maintain a certain portion of our dataset that remains private

Requirements - Contamination-Free Design - 2

Provided a question/task comes from a trusted source, we propose this formalism to accept it or not based on whether it overlaps (either textually or semantically) with tasks/questions already presented

Possible mathematical formalism for Contamination Detection

$$\text{Contamination\_Score} = \alpha \cdot \text{NGram\_Overlap} + \beta \cdot \text{Semantic\_Similarity} + \gamma \cdot (1 - \text{Temp\_Validity})$$

- **NGram_Overlap**: Measures direct textual similarity using 3,4,5-gram analysis to detect copied or closely paraphrased content
- **Semantic_Similarity**: Captures conceptual overlap using embedding-based methods to identify semantically equivalent content even when surface text differs
- **Temporal_Validity**: Binary indicator (1 if creation_date $>$ training_cutoff, 0 otherwise)
- **Weights (α, β, γ)**: Relative importance of each component - to be determined
- **Acceptance Threshold**: Contamination_Score cutoff value - to be established

Requirements - Comprehensive Reasoning Assessment

In light of these limitations, we must design a benchmark that assess the following abilities or types of reasoning

Generalization Testing Framework

Assessment Method:

- **Language Puzzles**: Crosswords with cultural clues, anagrams using Sanskrit-derived terms, traditional riddles (Paheli)
- **Evaluation**: Exact match scoring for puzzle solutions
- **Metric**: $P(\text{correct}|\text{novel combination})$ where novel combinations $\notin$ training examples

Open-Ended Reasoning

Assessment Method:

- **Multi-step logical inference**: Chain-of-thought reasoning in cultural contexts requiring 3+ reasoning steps
- **Causal reasoning**: Analyze cause-effect relationships in Indian scenarios (e.g., Green Revolution impacts)
- **Evaluation**: Human expert scoring on logical validity and cultural appropriateness

Requirements - Comprehensive Reasoning Assessment - 2

In light of these limitations, we must design a benchmark that assess the following abilities or types of reasoning

Toxic Penetration Testing

Assessment Method:

- **Cultural context sensitivity**: Test same lexical items across regions (e.g., surnames offensive in some regions)
- **Finer classification**: Moving beyond binary toxic/non-toxic judgment
- **Adversarial robustness**: Resistance to prompt injection and cultural manipulation attempts

Analytical Reasoning

Assessment Method:

- **Data interpretation**: Census analysis, economic indicators, social media trend analysis requiring statistical inference
- **Mathematical reasoning**: Festival-economic correlations, optimization problems specific to Indian events

Requirements - Comprehensive Reasoning Assessment - 3

In light of these limitations, we must design a benchmark that assess the following abilities or types of reasoning

Creative Reasoning

Assessment Method:

- **Story generation**: Human-AI discrimination task with 200-500 word stories in Indian cultural contexts

Temporal-Spatial Integration

Assessment Method:

- **Planning agent tasks**: Multi-step scenarios combining time and location constraints (Char Dham Yatra, festival logistics)
- **Event sequencing**: Historical timelines with geographical considerations
- **Resource coordination**: Agricultural cycles across different regions and seasons

Task Design – Generalization Testing with Language Puzzles

To test the generalization capabilities of LLMs, we can give them the following tasks/puzzles to solve

Crossword Puzzles

Design Specifications:

- Cultural clues: Indian festivals, personalities, concepts
- Script diversity: Devanagari, Tamil, Arabic, Gurmukhi
- Complexity levels: Bigger grids

Example:

```
Clue: "Festival of lights in Hindi (6 letters)"
Answer:  (Diwali)
```

Word Search Puzzles

- Hidden cultural terms: we put only indian concepts
- Multi-directional search: instead of the traditional 2D word search, we can complexify and create a 3D word search.

Task Design – Generalization Testing with Language Puzzles - 2

To test the generalization capabilities of LLMs, we can give them the following tasks/puzzles to solve

- Anagrams
- Riddles (Paheli)
- Other puzzles (Aksharit for example)

Task Design - Creative Reasoning and Planning

For creative reasoning, we can ask LLMs to produce stories that we submit to a simple Turing test.

Story Discrimination Task

Dataset Composition:

- **Human Authors (50%):**
- **LLM Generated (50%):**

Parameters:

- Length: 200-500 words
- Themes: Indian cultural contexts, universal themes
- Quality Control: $\kappa \geq 0.8$ inter-annotator agreement

Temporal-Spatial Planning

Scenario Categories:

1. Single-city navigation (Mumbai local trains)
2. Multi-state travel (Char Dham Yatra)
3. Festival coordination (Durga Puja logistics)
4. Agricultural planning (Rabi-Kharif cycles)

Evaluation Criteria:

- Feasibility, efficiency, appropriateness

Task Design - Open-Ended Reasoning

To test open-ended reasoning, we can survey many-subtypes of reasoning

Open-Ended Reasoning Tasks Questions

Causal Chain Reasoning

- "What would be the potential effects of implementing UPI payments in rural markets across different Indian states? Consider factors of infrastructure, literacy, and cultural adoption

Analogical Reasoning:

- **Cross-cultural analogies**: "If Diwali:Light::Holi:?, explain the underlying cultural pattern and apply this reasoning to other Indian festivals"

Creative Problem-Solving:

- **Language preservation**: "Propose methods to preserve endangered tribal languages while enabling community integration with mainstream society"

Task Design - Analytical Framework

Analytical Reasoning Framework

Data Interpretation:

- **Economic indicators**: "Analyze the correlation between monsoon patterns and agricultural GDP across different regions. What policy recommendations emerge?"

Logical Argument Construction:

- **Historical analysis**: "Construct logical arguments about the impact of British colonial education policies on contemporary Indian education systems"

Mathematical Reasoning:

- **Statistical inference**: "Calculate and interpret correlation between festival dates and regional economic activity patterns"

Dataset Derivation

Phase 1: Content Sources and Expert Network

Content to be Scraped:

- **For Planning Tasks**: Recent news articles about infrastructure projects, festival organization reports, agricultural planning documents
- **For Analytical Reasoning**: Contemporary census reports, economic surveys, social media datasets with temporal stamps
- **For Creative Reasoning**: Human Stories of 200-500 word by professional writers and native speakers

Ideally if we can find

- Linguists (Regional specialists) to help create crossword clues, anagrams, riddles
- Cultural historians to help design culturally authentic puzzle themes and story contexts
- Educational professionals to help develop difficulty-calibrated analytical reasoning questions
- Literature scholars to help write human-authored stories for discrimination tasks

Dataset Derivation - 2

Phase 2: Original Content Creation

What Needs to Be Generated:

- **Language Puzzles**: Crosswords, anagrams, riddles with cultural clues
- **Planning Scenarios**: Multi-step temporal-spatial coordination tasks
- **Reasoning Questions**: Open-ended analytical and creative problem-solving prompts

Quality Assurance:

- 100% human-generated base materials
- Native speaker validation

Phase 3: Validation Pipeline

For All Content Types:

1. Linguistic accuracy review
2. Cultural appropriateness assessment
3. Difficulty calibration per task type
4. Contamination verification using proposed detection algorithm
5. Inter-language consistency across scripts

Task-Specific Validation:

- **Puzzles**: Solvability verification
- **Stories**: Authenticity and quality
- **Planning**: Real-world feasibility

Scoring Mechanisms

These are scoring mechanism proposed for some of the tasks outlined above. Some others still need a scoring mechanism. These are mainly the open ended and analytical reasoning questions.

Exact Match Scoring (Puzzles)

```
def crossword_scoring(solution, reference):
    cell_accuracy = correct_cells / total_cells
    word_completion = completed_words / total_words
    cultural_score = evaluate_cultural_context(
        solution, reference)

    final_score = (0.4 * cell_accuracy +
                   0.4 * word_completion +
                   0.2 * cultural_score)
    return final_score
```

Pass@k Scoring (Planning)

```
criteria_weights = {
    'feasibility': 0.3,
    'efficiency': 0.25,
    'cultural_appropriateness': 0.25,
    'completeness': 0.2
}

# Pass@k: At least one plan above threshold
success_threshold = 0.7
return any(score >= success_threshold
           for score in plan_scores)
```

Scoring Mechanisms - 2

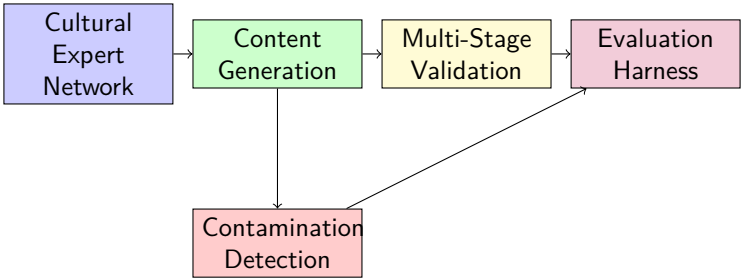
Binary Classification (Stories)

We ask humans to evaluate whether a story is human or LLM generated. Then we rely on how their accuracy deviates from random 50% accuracy to assess whether the stories are credible.

Performance Requirements

- **Contamination Resistance**: >95% verified clean
- **Open-ended Reasoning**: 70%+ human correlation
- **Cultural Toxicity**: 85%+ F1 multi-class
- **Temporal-Spatial**: 60%+ feasibility score

Implementation Overview



Key Innovations

Key Innovations

- **Contamination-Free**: Temporal boundaries + semantic verification
- **Culturally-Aware**: Regional sensitivity + expert validation
- **Reasoning-Comprehensive**: Missing reasoning types addressed
- **Evaluation-Rigorous**: Multi-criteria scoring with cultural context