Buddhi-Pragati: Addressing Gaps in LLM Reasoning Benchmarks in Indic Languages

by

Email:
Repository: https://anonymous.4open.science/r/buddhi-pragati-321E/

September 2025

Abstract

Current benchmarks for evaluating Large Language Model (LLM) reasoning capabilities in Indic languages predominantly assess knowledge retrieval and linguistic competence tasks (question answering, summarization, text translation, etc.). They also rely on culturally irrelevant or neutral content. This means that the reasoning abilities of LLMs in such low resource language remain largely unexplored. This work introduces Buddhi-Pragati, a novel benchmark framework that employs crossword puzzles to evaluate rule-based reasoning in 19 Indic languages. We present three key contributions: (1) the identification of crossword puzzles as memorization-resistant evaluation instruments for low-resource language reasoning, (2) a corpus of 120,000 culturally-authentic clue-answer pairs scored for Indian contextual relevance using a novel embedding-based approach, and (3) a dataset of 4,750 crossword puzzles generated through memetic algorithms optimized for constraint density and cultural authenticity. Our framework also implements seven experimental configurations across 20+ LLMs, evaluating performance through word accuracy, letter accuracy, and intersection consistency metrics. By grounding our clues-answers corpus in culturally-relevant content while maintaining resistance to memorization, Buddhi-Pragati addresses critical gaps in current Indic language benchmarking paradigms. The complete datasets and evaluation harness are publicly available to facilitate reproducible research in multilingual reasoning assessment. [1] [2] [3]

Contents

[1] https://huggingface.co/datasets/selim-b-kh/buddhi-pragati
[2] https://huggingface.co/datasets/selim-b-kh/buddhi-pragati-puzzles
[3] https://anonymous.4open.science/r/buddhi-pragati-321E/

1 Introduction

The evaluation of Large Language Models (LLMs) has evolved from task-specific assessments to increasingly sophisticated benchmarks that attempt to measure genuine reasoning capabilities. Despite serving over 1.4 billion speakers collectively, Indic languages remain critically underrepresented in both LLM training corpora and evaluation frameworks, typically comprising less than 5% of training data even in multilingual models [11, 21]. This representation gap justifies the need for language specific and culture specific benchmarks.

Existing Indic language benchmarks predominantly adopt translation-based approaches, adapting established English benchmarks such as MMLU [9] and SuperGLUE [25] to Indian languages. While these efforts provide standardized evaluation metrics, they fail to capture two critical dimensions of reasoning assessment. First, they remain vulnerable to memorization-based solutions, where models achieve high scores through pattern matching rather than genuine reasoning. Second, they lack cultural authenticity, grounding evaluation in Western conceptual frameworks that may not reflect Indian reasoning patterns or knowledge systems.

Recent theoretical advances in intelligence measurement, particularly Chollet's formalization of intelligence as "skill-acquisition efficiency over a scope of tasks," suggest that effective benchmarks must evaluate compositional reasoning over novel situations rather than accumulated knowledge from training [4]. This paradigm shift motivates us to explore crossword puzzles as potential intelligence and reasoning proxies. Crosswords inherently require multi-constraint reasoning, where solvers must simultaneously satisfy lexical, semantic, and spatial constraints while adapting to partially-solved states. The combinatorial explosion of possible puzzle configurations makes exhaustive memorization impractical, which means that high performance on such benchmarks genuine reasoning capabilities.

This work introduces **Buddhi-Pragati**, a comprehensive framework for evaluating LLM reasoning capabilities in Indic languages through culturally-authentic crossword puzzles. Our approach addresses three fundamental challenges in multilingual reasoning assessment. First, we develop a novel corpus construction methodology that extracts and scores clue-answer pairs for cultural relevance using embedding-based similarity metrics. Second, we implement memetic algorithms for puzzle generation that optimize for both structural quality and cultural authenticity. Third, we establish experimental protocols that systematically evaluate reasoning performance across complexity levels, prompting strategies, and model architectures.

The primary contributions of this research include:

(1) a theoretical justification for crossword-based reasoning evaluation in low-resource languages

(2) a corpus of 120,000 clue-answer pairs across 19 Indic languages with cultural context scoring,

(3) 4,750 generated crossword puzzles optimized for constraint density and cultural relevance

(4) an extensible evaluation framework supporting multiple experimental configurations and model architectures.

By grounding reasoning assessment in culturally-authentic content while maintaining memorization resistance, Buddhi-Pragati establishes a new paradigm for evaluating genuine reasoning capabilities in underrepresented languages.

2 Related Work

2.1 Global LLM Benchmarking Paradigms

Historically, there have been two competing paradigms when it comes to LLM evaluation [4]. The first sees intelligence as a collection of task-specific skills, leading to benchmarks that focus on individual abilities. In language benchmarks, this has given evaluations on question answering, text translation, sentiment analysis, and summarization [25, 24]. For instance, the highly praised MMLU benchmark [9] evaluates models across 57 subjects with multiple-choice questions, treating intelligence as the aggregation of domain-specific competencies. Similarly, SuperGLUE [24] extended GLUE [25] with more challenging tasks but maintained the same fundamental assumption that intelligence equals the sum of discrete skills. Sadly, this approach has led to ever harder benchmarks needed to be released more and more frequently to be able to discriminate between models, as the drastic increase in released models' sizes and compute units used in training has made models better and better memorizers [10]. This has also meant that human performance is no longer easily comparable to LLM performance in these benchmarks, as they require high mastery level of a certain skill. Put plainly, these benchmarks either exclude human performance or treat it as the "performance of the best human". Furthermore, as [4] argues, at least for much of these benchmarks, a model could solve them if it had access to unlimited training, without it impacting its reasoning abilities.

This is why some contemporary benchmarks [4, 18] have shifted towards measuring intelligence through a different lens: the ability of models to reason over unknowns through compositional rule-based reasoning and environmental adaptation, which we will come back to later. In other words, this second paradigm treats intelligence as a general learning ability, with [4] formalizing it as "a measure of skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty". In recent work, ARC-AGI [3] exemplifies this shift as it evaluates what it calls "fluid" intelligence through novel pattern recognition tasks that resist memorization-based solutions.

2.2 Indic Languages Benchmarks

Not all languages are equal when it comes to LLM performance with languages dominating the training corpora of models, being naturally advantaged compared to others. Indic languages, despite being spoken by more than 1.4 billion people with each gathering millions to hundreds of millions of speakers, remain significantly underrepresented in LLM training corpora. They form no more than 8% of the content of training corpora [27, 21] designed specifically for multilingual abilities and around 0.5 to 5% of normal corpora. This representation and resources gap [11] has motivated the development of specialized evaluation frameworks for Indic language capabilities.

Efforts focused on translating established English benchmarks. For example, IndicMMLU-Pro [12] adapts MMLU's multiple-choice format across 14 domains for 9 Indic languages. While such translation-based approaches leverage existing quality control mechanisms in other benchmarks, they often fail to capture Indian concepts and ground the evaluation of LLMs in Indic languages in a non-Indian framework. We will come back to this later.

More recent benchmarks have emphasized cultural integration and contextual relevance. MILU represents a significant advancement by incorporating content from national, regional and state-level examinations across 11 Indic languages, spanning 42 subjects in 8 domains [23]. This approach grounds evaluation in authentic Indian educational contexts, testing models' familiarity with local knowledge systems including regional history, cultural practices, and legal frameworks.

IndicGenBench extends evaluation beyond knowledge assessment to generation capabilities, covering 29 languages across four tasks: cross-lingual summarization, machine translation, multilingual question answering, and cross-lingual question answering [19].

Despite these advances, current Indic benchmarks remain predominantly focused on knowledge retrieval and basic language understanding tasks. IndicNLG and IndicNLU evaluate natural language generation and understanding capabilities respectively [14, 6], but these assessments do not test reasoning or problem-solving abilities in Indic languages. Even MILU, while culturally grounded, essentially tests accumulated knowledge rather than the ability to reason over novel situations using cultural priors.

2.3 Key Limitations

In light of this, we identify the following limitations to the current reasoning benchmarks for LLMs in Indic languages:

1. **Vulnerability to memorization and lack of assessment of broad generalization abilities**.While benchmarks like MILU and IndicGenBench test accumulated knowledge and linguistic competence, they fail to evaluate compositional rule-based reasoning and environmental adaptation.
2. **Contamination prevention**. Contamination presents a pervasive threat to evaluation validity. MEGAVERSE's systematic analysis reveals extensive contamination across major multilingual models, with particularly concerning evidence for low-resource languages where detection methods are less robust [1]
3. **Lack of cultural authenticity**. Cultural authenticity remains compromised by over-reliance on translation-based approaches, which may fail to capture reasoning patterns specific to Indian cultural and educational contexts

2.4 Indic Reasoning Benchmarks

To address these limitations, we propose and explore the use of crossword puzzles as they require genuine multi-rule based reasoning, rather than shallow pattern-matching. To solve a crossword puzzle, models need to compose with many constraints stemming from the spatial structure of a grid (which cells are not all accessible) and the solving of a clue (which can be more or less cryptic), all while being able to reason through how previously solved clues impact new word placements.

As for the memorization-safe requirement, the memorization-resistance properties of crosswords stem from their inherent structure. Unlike fixed question-answer pairs, crossword clues are contextually ambiguous and require inference based on intersecting constraints. The combinatorial increase of possible puzzle configurations makes exhaustive memorization impractical.

CrosswordBench's [16] has already established the viability of crossword-based assessments in English and Chinese and showed that even advanced LLMs struggle with these puzzles, with performances barely reaching 50% accuracy. As far as we know, no existing Indic benchmark exploits these properties for reasoning evaluation. This gap is particularly significant given the cultural and linguistic diversity of Indian contexts, which does require different cultural priors and reasoning than those captured by English-centric evaluations.

The main contributions of this paper reside in the following:

1. The identification of a new path to benchmark LLM reasoning in low-resource languages

2. The curation of a dataset of nearly 120000 clue-answer pairs across 19 Indic languages scored on their suitability for crossword puzzles and relatability to Indian context
3. The curation of a dataset of 4750 crossword puzzles across 19 indic languages, built from the aforementioned corpus using a memetic generator algorithm inspired by [7] and designed to achieve a target density and to guarantee intersection words
4. Package that lets you perform 7 experiments on these datasets using 20+ LLMs (open and closed source) and evaluate them using 3 metrics

3 Methodology

3.1 Clue-answer pairs dataset curation

3.1.1 Identifying suitable sources

Before generating puzzles, we need to curate a corpus of clue-answer pairs. To do so, we must ensure to select suitable sources from which to draw our pairs. Concretely we needed to:

- Target sources that provide straightforward clue-answer mappings
- Eliminate the sources that require external context beyond linguistic knowledge such as sources where the clue component would be an MCQ of the following format: "Among the following options, choose ...."
- Eliminate the cryptic crosswords given that they would be too complex for a first crossword-based benchmark and would involve too many composition of rules.
- Ensure thematic diversity to prevent generating puzzles in the same topic, which would compromise puzzle variety

After comprehensive analysis, we identified four complementary data sources that collectively satisfy these requirements and span the required linguistic and thematic diversity:

- **MILU** [22]: A multi-task Indic language understanding dataset containing multiple-choice questions across 11 languages, filtered to extract direct question-answer pairs suitable for crossword clues.
- **Bhasha-wiki** [20]: Wikipedia articles in Indic languages processed through named entity recognition to extract entity-description pairs, providing encyclopedic knowledge coverage.
- **IndicWikiBio** [15]: Biographical entries from Wikipedia containing structured infoboxes, enabling extraction of name-based clues from biographical summaries.
- **IndoWordNet** [2]: A multilingual wordnet for Indian languages providing dictionary definitions that can be transformed into definitional crossword clues.

Each data source then goes through processing to extract crossword-suitable content while maintaining consistency across the multilingual corpus. This processing is detailed here.

MILU Processing: Multiple-choice questions undergo contextual pattern detection using language-specific regex patterns to exclude questions requiring option selection. Latin script filtering ensures script consistency in Indic language subsets, while enhanced single-word answer validation prevents multi-token responses unsuitable for crossword grids.

Bhasha-Wiki Processing: Named entity recognition using ai4bharat/IndicNER extracts entities from Wikipedia articles, with first-sentence extraction providing concise clues. The pipeline implements dual-mode processing for English versus Indic content, ensuring appropriate entity-description pairing while filtering Latin script entities from predominantly Indic texts.

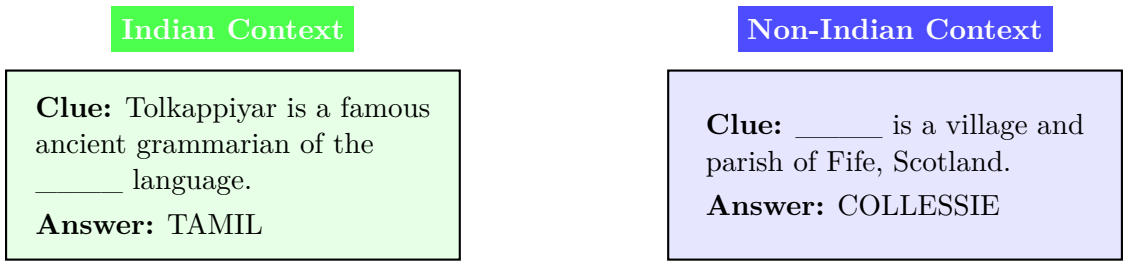
IndicWikiBio Processing: Biographical infoboxes undergo systematic name extraction with comprehensive name removal from summary text to prevent answer leakage. The system processes sequential name fields (name_1, name_2, etc.) while cleaning HTML artifacts and phonetic pronunciations to ensure crossword-appropriate clues.

IndoWordNet Processing: Dictionary definitions are processed to create definitional clues with target word removal to prevent answer revelation. The pipeline ensures single-word target validation while maintaining definitional clarity suitable for crossword solving.

Universal Quality Control: All sources implement answer-in-clue prevention using Unicode-aware validation, length constraints (2-12 characters for answers, 10-500 characters for clues), character diversity requirements, and comprehensive rejection tracking across three logging tiers.

3.1.2 Defining and Computing the Context Score

Then, as we mentioned in section 2, we have to assess "how indian" our pairs are. This is what distinguishes our benchmark from translation-based approaches. We define "Indian context" as clue-answer pairs involving cultural, political, geographical, or historical references specific to the Indian subcontinent. Figure 1 illustrates the distinction between culturally-relevant and generic entries.



(a) Culturally-relevant entry

(b) Generic entry

Figure 1: Examples of context scoring from the Buddhi-Pragati dataset (`https://huggingface.co/datasets/selim-b-kh/buddhi-pragati`)

Our context scoring methodology uses embedding-based similarity computation augmented with keyword matching. Initial experiments using single-word embeddings ("India") produced inconsistent results for what we would classify as indian pairs, motivating us to develop the following corpus-based approach. The reference embedding construction process begins with a manually-curated seed corpus o Indian concepts spanning geography ("Ganges", "Western Ghats"), culture ("Diwali", "Kathakali"), history ("Mughal Empire", "Independence Movement"), and contemporary references ("ISRO", "IIT"). We expand this seed corpus using IndoWordNet's semantic relationships, specifically extracting hyponyms (more specific concepts) and meronyms (part-whole relationships).

Then, we use L3Cube-IndicSBERT [5] to compute the reference embedding for its 11 supported languages, utilizing its cross-lingual alignment properties to maintain consistency across languages. For remaining languages, we employ sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [17], which provides adequate multilingual coverage despite lower Indic-specific optimization.

The final context score combines cosine similarity with keyword matching:

$$\text{Context Score} = \max(1, \alpha \cdot \cos(\vec{e}_{\text{entry}}, \vec{e}_{\text{ref}}) + \beta \cdot |\text{keywords} \cap \text{entry}|)$$

where $\vec{e}_{\text{ref}}$ represents the average embedding of the Indian corpus, α and β are tunable weights.

3.2 Puzzles dataset curation

Inspiring ourselves from [7], our generation pipeline separates the crossword creation process into two distinct phases: grid structure optimization and constraint satisfaction through word placement.

Phase 1: Grid Structure Generation employs a memetic algorithm that combines evolutionary search with local optimization to generate high-quality crossword masks. The fitness function is calculated using multiple quality criteria: density of the grid, word length balance to make sure the puzzles are solvable but complex, clustering minimization to ensure even distribution of definition fields, and intersection optimization maximize constraint-based reasoning opportunities.

Phase 2: Constraint Satisfaction and Word Placement implements a backtracking algorithm enhanced with constraint propagation to fill generated grids with culturally appropriate clue-answer pairs. The placement algorithm prioritizes entries with higher cultural context scores, ensuring that puzzles maintain cultural authenticity while satisfying intersection constraints. The system implements intelligent ordering heuristics that prioritize word placement based on constraint density and cultural relevance scores.

3.2.1 Memetic Optimization Framework

Our memetic algorithm implementation addresses the limitations of pure genetic approaches identified in the crossword manufacturing literature [7] by incorporating hill-climbing local search as a repair mechanism for evolutionarily generated solutions.

The algorithm maintains a population of crossword grids, applying crossover operations followed by localized repair procedures. The repair mechanism addresses constraint violations introduced during crossover, particularly at grid boundaries where different parents' structural patterns may conflict. This hybrid approach combines the exploration capabilities of evolutionary search with the exploitation effectiveness of local optimization.

3.2.2 Cultural Context Integration

As mentioned above, to ensure cultural authenticity in generated puzzles, our system implements context-aware word selection during the constraint satisfaction phase. The integration mechanism implements a weighted scoring system that balances cultural authenticity with puzzle solvability. High cultural context scores receive placement priority, but the system maintains flexibility to select lower-scoring alternatives when constraint satisfaction requires specific word characteristics (length, letter patterns). This ensures that generated puzzles maintain cultural coherence while remaining structurally sound.

Grid density optimization targets 75% cell occupancy, aligning with standard crossword puzzles while providing sufficient constraint density to challenge reasoning capabilities. The memetic algorithm incorporates density constraints into its fitness function, penalizing grids that deviate significantly from target occupancy levels.

The system generates puzzles across multiple grid sizes (7x7, 10x10, 15x15, 20x20, 25x25) to provide difficulty scaling for evaluation purposes. Larger grids increase compositional complexity.

Algorithm 1 Memetic Algorithm for Crossword Grid Generation

Require: Population size N, grid dimensions (w, h), cultural corpus C
Ensure: High-quality crossword grid G^*
Initialize population $P = \{G_1, G_2, \ldots, G_N\}$ randomly
Evaluate fitness $f(G_i)$ for all $G_i \in P$
while termination criteria not met **do**
  Select parents G_p, G_q using tournament selection
  Generate offspring G_o via crossover operation
  if G_o violates validity constraints **then**
    Apply hill-climbing repair: $G_o \leftarrow \text{HillClimb}(G_o)$
  end if
  Evaluate fitness $f(G_o)$
  Apply mutation with probability p_m
  if mutation applied **then**
    $G_o \leftarrow \text{CentralizedMutation}(G_o)$
    Re-evaluate fitness $f(G_o)$
  end if
  Update population using $(\mu + \lambda)$ selection
end while
Select best grid $G^* = \arg\max_{G \in P} f(G)$
Fill grid using cultural context-weighted backtracking
return Complete crossword puzzle with cultural authenticity

4 Experiments

4.1 Metrics

We adopt the evaluation metrics from CrosswordBench [16] to assess model performance across multiple dimensions. Following their hypothesis regarding directional performance asymmetry (ie LLMs supposedly perform better for horizontal words), we decompose metrics to evaluate horizontal (across) and vertical (down) accuracy independently.

The evaluation framework employs three primary metrics:

Word Accuracy measures the proportion of correctly completed words:

$$\text{WA} = \frac{1}{|W|} \sum_{w \in W} \mathbb{I}[\text{predicted}(w) = \text{reference}(w)]$$

where W represents the set of all words in the puzzle. We compute directional variants:

$$\text{WA}_{\text{across}} = \frac{1}{|W_{\text{across}}|} \sum_{w \in W_{\text{across}}} \mathbb{I}[\text{predicted}(w) = \text{reference}(w)]$$

$$\text{WA}_{\text{down}} = \frac{1}{|W_{\text{down}}|} \sum_{w \in W_{\text{down}}} \mathbb{I}[\text{predicted}(w) = \text{reference}(w)]$$

Letter Accuracy quantifies character-level correctness:

$$\text{LA} = \frac{1}{|L|} \sum_{l \in L} \mathbb{I}[\text{predicted}(l) = \text{reference}(l)]$$

where L denotes all letter positions in the grid.

Intersection Consistency Rate measures constraint satisfaction at word intersections:

$$\text{ICR} = \frac{1}{|I|} \sum_{i \in I} \mathbb{I}[\text{across}(i) = \text{down}(i)]$$

where I represents intersection points between horizontal and vertical words.

4.2 Experimental Configurations

Once we have defined all this, we can start running experiments. 'Buddhi-pragati' involves a framework for 10 different experiments, investigating the effect on performance of linguistic factors, prompting strategies, and model architectures. We detail them below:

4.2.1 Prompting Strategy Experiments

Experiment 0 (Master experiment compares the performance across all puzzles for all models in all languages. This, like the following experiments is set up using the baseline configuration and is the only one running on the entirety of our parameter space. In fact the experiments 8 to 10 are just re-reading of this one.

Experiment 1 (Language Family Analysis) compares the performance across Dravidian languages (Tamil, Telugu, Kannada, Malayalam) versus Indo-Aryan languages (Hindi, Bengali, Gujarati, Marathi, Punjabi, Urdu, Odia) to identify systematic linguistic processing differences.

Experiment 2 (Few-Shot Learning Evaluation) compares zero-shot, one-shot, and few-shot prompting to assess in-context learning capabilities. The objective of this experience is to test the conclusions of [13] Few-shot examples are randomly sampled from the same language but varied grid sizes to minimize the risk of providing the LLMs with clues they might already have seen, show that the solving logic spans many grid sizes and maintain linguistical coherence in the prompt.

Experiment 3 (Batch Size Effects) evaluates single-puzzle versus 10-puzzle versus all-puzzle batch processing to determine whether models learn during evaluation sessions. In the case where this experiment is bundled with experiment 2, few-shot examples are provided as batch headers rather than per-puzzle to maintain consistent evaluation conditions and avoid giving too many examples to the model.

Experiment 4 (Chain-of-Thought Reasoning) compares explicit outputted reasoning before the answer generation versus direct answer generation to assess the impact of intermediate reasoning steps on crossword solving accuracy

Experiment 5 (Reasoning Effort Analysis) varies token limits (500 for low-effort, 1000 for normal, 2000 for high-effort) to determine the relationship between computational resources and solving performance.

Experiment 6 (Self-Reflection Capabilities) implements iterative solving where models receive feedback on incorrect placements and attempt corrections, testing how well do models recover from their errors and adapt their reasoning.

Experiment 7 (Cross-Linguistic Prompting) compares same-language prompting versus English-prompt with target-language puzzle to evaluate cross-linguistic instruction following capabilities. This experiment is partially set up. For it to be completely set up, we would need to translate all prompt into every single of the 19 languages, which ideally would require using some model. One could use IndicTrans2, which performance presented in [8] and usage in [22] make it promising.

4.2.2 Model Architecture Experiments

Experiment 8 (Specialization Analysis) compares Indic fine-tuned models against general multilingual models to quantify the impact of domain-specific training on reasoning performance. The key insights we would like to measure is whether Indic Fine-Tuned models reason better over Indic languages than general multilingual models. The obvious answer would be yes but as seen from [22], this is not straightforward.

Experiment 9 (Reasoning Architecture Comparison) evaluates reasoning-optimized models (e.g., GPT-o1, Claude-3.5) against standard language models to assess architectural innovations for complex reasoning tasks.

Experiment 10 (Efficiency Normalization): Performance analysis normalized by computational cost (cost/token) and resource utilization (token usage) to identify optimal efficiency-performance trade-offs.

4.2.3 Implementation details

To implement the experiments detailed above, our framework uses modular prompt templates so that the user can combine experiments together and evaluate a combined prompting strategy. Then, we have an experiments orchestrator that separates the prompting methodology experiments (1-7) from model architecture experiments (8-10). We then define a baseline configuration that all experiments start from.

Baseline Configuration: The default experimental configuration is the following one: zero-shot prompting, single-puzzle evaluation batches, direct answer generation (no chain-of-thought), normal reasoning effort (1000 tokens), no self-reflection, and monolingual prompting. All experimental variations modify precisely one parameter from this baseline to ensure controlled comparisons.

Then experiments 1 to 7 only run on a part of our parameter space which is the 'priority' one:

Priority Evaluation Set: The priority subsets for intensive analysis: grid sizes 7×7, 15×15, and 25×25; ten core languages spanning Indo-Aryan (Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Urdu) and Dravidian (Kannada, Malayalam, Tamil) families plus English; and a curated model set balancing reasoning capabilities, multilingual support, and computational efficiency.

At the end, all experimental results are systematically logged in structured JSON format, so that we can compare results cross-experiment.

5 Limitations

Our work presents several important limitations that we must acknowledge for proper interpretation of the methodology proposed by 'Buddhi-Pragati'.

- **Experimental Validation Gap**: This work lays the theoretical framework, datasets, evaluation infrastructure, and experimental configurations for a study of the abilities of LLMs on crossword puzzles. However, the actual performance results across the proposed 10 experimental conditions remain to be established in subsequent research
- **Dataset Contamination Risk**: As highlighted by [1, 26], our curated datasets may overlap with model training corpora, potentially inflating performance estimates. Our methodology does not implement contamination detection mechanisms or even a private

dataset of puzzles as in [4], which represents a significant threat to validity in multilingual LLM evaluation.

- **Human Evaluation Absence**: The crossword puzzle quality and, especially the validity and solvability of the clues rely entirely on computational metrics without human validation from native speakers of the target languages. This limitation is particularly critical, especially given that the author does not speak any of the targeted Indic language nor can he read any. To rectify this, we could apply an experimental flow similar to that in [12] or [28]

5.1 Future Research Directions

The limitations identified above suggest several avenues to extend this work:

- **Alternative Puzzle Formats**: We could expand beyond crosswords. Other constraint-satisfaction puzzles (Sudoku variants, word search, anagrams) could provide complementary insights into multilingual reasoning capabilities while maintaining the cultural context evaluation framework. We could also curate puzzles that blend other reasoning constrain-based rules (for example a puzzle requiring to manage temporal and spatial constraints)
- **Enhanced Cultural Scoring**: Investigating alternative approaches to cultural relevance assessment, including using corpora validated by a native community, or hierarchical cultural taxonomy systems, could improve the precision and cultural sensitivity of context evaluation.

6 Conclusion

This work introduces Buddhi-Pragati, a novel evaluation framework that addresses fundamental gaps in multilingual reasoning assessment for Large Language Models through crossword puzzle solving across 19 Indic languages. By integrating constraint-satisfaction problems with cultural context awareness, we establish a new paradigm for evaluating genuine reasoning capabilities that extends beyond surface-level pattern recognition to require authentic linguistic and cultural understanding.

Our primary contribution lies in the comprehensive dataset curation methodology that systematically processes four complementary sources (MILU, IndicWikiBio, IndoWordNet, Bhasha-Wiki) to generate 120,000 culturally-grounded clue-answer pairs. The novel cultural context scoring system, leveraging IndoWordNet's semantic relationships and multi-tier embedding similarity, provides robust assessment of Indian cultural relevance while maintaining scalability across the entire linguistic spectrum. This approach successfully addresses the critical limitation of culturally-neutral content in existing multilingual benchmarks.

The mathematical formalization of evaluation metrics, extending CrosswordBench's directional analysis to multilingual contexts, enables precise quantification of reasoning performance across Word Coverage Rate (WCR), Letter Coverage Rate (LCR), and Intersection Consistency Rate (ICR). These metrics, combined with our systematic experimental framework examining ten distinct conditions, provide comprehensive insights into the interplay between linguistic families, prompting strategies, and model architectures in constraint-satisfaction reasoning tasks.

The technical innovation of memetic algorithm-based puzzle generation, optimized for cultural authenticity and linguistic diversity, demonstrates the feasibility of creating computationally challenging evaluation environments that resist contamination while maintaining cultural specificity. The framework's modular architecture and comprehensive experimental design establish a

replicable methodology for extending reasoning evaluation to additional low-resource languages and alternative constraint-satisfaction problems.

While this work presents theoretical frameworks and evaluation infrastructure without empirical results, the contributions establish essential foundations for advancing multilingual reasoning assessment. The identification of systematic limitations—including contamination risks, human validation absence, and cultural context scope—provides clear directions for future research while acknowledging the inherent challenges in creating authentic multilingual evaluation environments.

The broader implications extend beyond technical evaluation to address fundamental questions about cultural authenticity in AI assessment. By demonstrating how constraint-satisfaction problems can incorporate cultural specificity while maintaining computational rigor, this work contributes to the growing recognition that effective multilingual AI evaluation requires deep integration of linguistic, cultural, and computational considerations rather than simple translation or adaptation of English-centric methodologies.

Future applications of this framework promise to illuminate critical aspects of multilingual reasoning that remain underexplored in current literature. The systematic comparison of reasoning versus non-reasoning models, Indic-specialized versus general architectures, and cross-linguistic prompting strategies will provide essential insights for developing more culturally-aware and linguistically-diverse AI systems. The framework's extensibility ensures continued relevance as both model capabilities and evaluation needs evolve in the rapidly advancing landscape of multilingual artificial intelligence.

References

[1] Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. Megaverse: Benchmarking large language models across languages, modalities, models and tasks, 2024.

[2] Pushpak Bhattacharyya. IndoWordNet. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[3] Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2025.

[4] François Chollet. On the measure of intelligence, 2019.

[5] Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert, 2023.

[6] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages, 2023.

[7] Jakob Engel, Markus Holzer, Oliver Ruepp, and Frank Sehnke. On computer integrated rationalized crossword puzzle manufacturing. In Evangelos Kranakis, Danny Krizanc, and Flaminia Luccio, editors, *Fun with Algorithms*, pages 131–141, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[8] Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, 2023.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[11] Sankalp KJ, Vinija Jain, Sreyoshi Bhaduri, Tamoghna Roy, and Aman Chadha. Decoding the diversity: A review of the indic ai research landscape, 2024.

[12] Sankalp KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding, 2025.

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[14] Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages, 2022.

[15] Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages. 2022.

[16] Jixuan Leng, Chengsong Huang, Langlin Huang, Bill Yuchen Lin, William W. Cohen, Haohan Wang, and Jiaxin Huang. Crosswordbench: Evaluating the reasoning capabilities of llms and lvlms with controllable puzzle generation, 2025.

[17] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[18] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025.

[19] Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] Soket Labs Technology and Research Private Limited. Bhasha-wiki.

[21] Aatman Vaidya, Tarunima Prabhakar, Denny George, and Swair Shah. Analysis of indic language capabilities in llms, 2025.

[22] Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. Milu: A multi-task indic language understanding benchmark. *arXiv preprint arXiv: 2411.02538*, 2024.

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

[24] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020.

[25] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[26] Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024.

[27] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[28] Kamyar Zeinalipour, Mohamed Zaky Saad, Marco Maggini, and Marco Gori. From arabic text to puzzles: Llm-driven development of arabic educational crosswords, 2025.