TeluguEval: A Comprehensive Benchmark for Evaluating LLM Capabilities in Telugu

Revanth Gundam
IIIT Hyderabad
revanth.gundam@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Large Language Models (LLMs) excel on English reasoning tasks but falter on morphologically rich, low-resource languages such as Telugu, Tamil, and Kannada. We present TeluguEval, a human-curated reasoning benchmark created by translating GSM8K (math), Winogrande (commonsense), ARC (science), CaseHOLD (law), and Hendrycks Ethics into Telugu. We evaluate eight models spanning global (Llama-3.1-8B, Llama-2-7B, Qwen-8B, Gemma-7B, Gemini-2.0) and regional (Telugu-Llama2-7B, Indic-Gemma-7B, Sarvam-m-24B) systems. While extremely strong models such as Gemini and Sarvam-m largely retain performance in Telugu, most English-centric models suffer severe accuracy drops, often exceeding 30 to 40 points, particularly on mathematical and scientific reasoning. We further observe systematic failure modes including script sensitivity, option-selection bias, repetition loops, and unintended code-switching. Our results demonstrate that surface-level Telugu fluency does not imply robust reasoning capability, underscoring the need for Telugu-specific data, tokenization, and pretraining. TeluguEval provides a standardized testbed to drive progress on reasoning in low-resource Indian languages.

1 Introduction

Large language models based on the transformer architecture have led to great progress in NLP. Recent LLMs such as GPT-4, Claude-3, and Llama-3 perform well on a wide range of tasks, including multi-step reasoning, code generation, and creative writing. As a result, many traditional benchmarks have become less useful, and evaluation has shifted toward harder problems involving mathematical reasoning, logical consistency, and agentic behavior. Despite these advances, progress has not been even across languages. Most of the available training data, benchmarks, and model decisions continue to focus on English and a small number of high-resource European and East Asian languages, leaving many widely spoken languages underrepresented in both training and evaluation.

For speakers of many of the low-resource languages, the advantages or usefulness of the recent gains in LLMs remains quite limited. These languages also include those which are widely spoken but not heavily represented in good quality training data, such as Telugu. Telugu, with over 80 million speakers in India, is a Dravidian language with a long literary tradition. It is morphologically rich and agglutinative in structure. While there is some amount of Telugu text available online, most of it is not suitable for the current training paradigms of LLMs. As a result, Telugu users end up using multilingual models that use a fixed set of parameters across many languages. Such a setup usually leads to weaker performance, increased hallucinations and poor cultural alignment compared to languages like English. This highlights a gap between apparent multilingual coverage and real world reasoning capability.

1.1 The Limitations of Existing Indic Benchmarks

The current evaluation setup for Indic languages is not sufficient to measure the real capabilities of modern large language models. Early benchmarks such as IndicGLUE (Kakwani et al., 2020) made an important contribution by standardizing evaluation across 11 Indian languages, but they mainly focused on discriminative tasks like news classification and fill in the blank style questions. While these tasks were reasonable during the BERT era, they are now too easy for generative language models. In many cases, models can solve them through surface level pattern matching, without needing to perform multi step reasoning.

Link to the code and data.

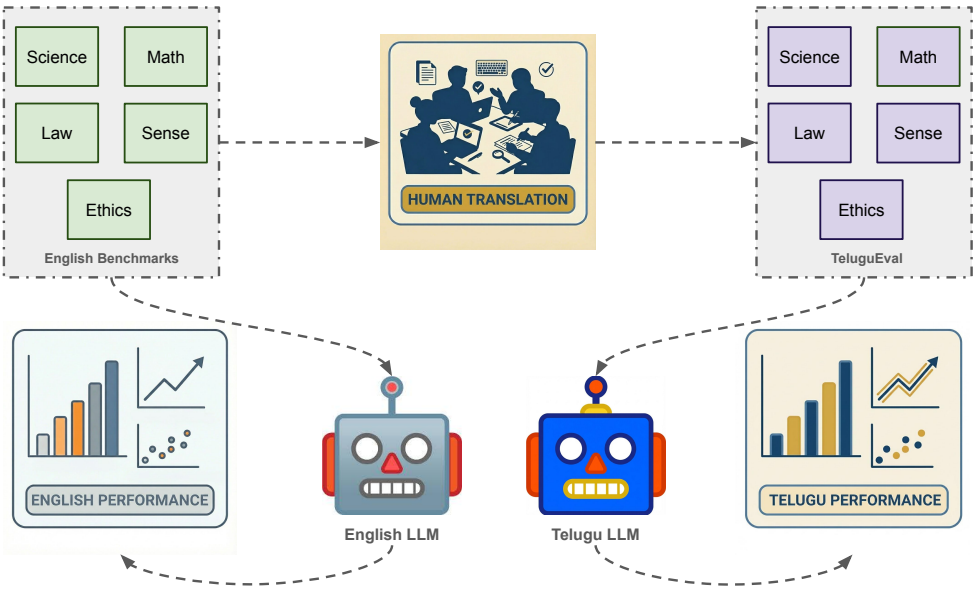

Figure 1: Overview of the TeluguEval construction and evaluation pipeline. We manually translated five gold-standard reasoning datasets into Telugu to create our comprehensive benchmark suite. This human-curated data is then used to evaluate and compare the performance of global English-centric LLMs against specialized regional models across both languages.

Much recent work such as IndicGENBENCH (Singh et al., 2024) has started to look at generative tasks like summarization, and Airavata (Gala et al., 2024) has introduced instruction following evaluations for Hindi. However, there is still a major gap when it comes to evaluating reasoning in Telugu. This includes the ability to break down a complex problem, follow a clear line of reasoning, and arrive at the correct answer. A model may translate a sentence accurately or classify a news article, but it is unclear whether it can solve a grade school math problem written in Telugu, understand a legal concept expressed in Telugu, or resolve a commonsense ambiguity that depends on local cultural context.

Many evaluations rely on machine translated versions of English benchmarks, which often introduce translation errors that break the intended reasoning task. For example, a math problem that depends on precise wording can become unclear or even unsolvable due to a single mistranslated word. This can lead to false failures, where a model is penalized because of bad data rather than poor reasoning. On the other hand, some translations simplify the original problem and remove important linguistic cues, leading to false successes where models appear to perform well without actually solving the intended task.

1.2 TeluguEval

To address these inherent gaps, we introduce TeluguEval, a comprehensive benchmark addressing diverse reasoning problems and domains in Telugu. These samples were manually translated by Telugu speakers as opposed to using noisy automated pipelines to maintain the quality of the data in Telugu. We have adapted and translated 5 standard and widely used datasets in English, GSM8k, Winogrande, ARC, CaseHOLD, and Ethics, into Telugu.

TeluguEval consists of 949 human-verified examples across five domains: Mathematics, Common-Sense, Science, Law, and Ethics. Specifically, each example is provided in both native Telugu script and a phonetic Romanized (Latin) version. We utilize this benchmark to evaluate eight widely used LLMs, ranging from global models (Llama-3.1-8B, Gemini-2.0 Flash) to region specific models (Sarvam-m-24B, Indic-Gemma-7B).

1.3 Contributions

In this paper, we make the following contributions:

- **TeluguEval Benchmark**: **We introduce *TeluguEval*, the first multi-domain, reasoning-focused benchmark for Telugu, comprising 949 human-verified examples across**

mathematics, commonsense reasoning, science, law, and ethics.

- **High-Quality Human Translation**: All benchmarks are expert-translated and native-speaker verified, avoiding common pitfalls of machine-translated evaluations and preserving the original reasoning complexity.
- **Comprehensive Model Study**: We benchmark eight global and regional LLMs under identical prompting conditions, revealing substantial performance gaps between English-centric and Telugu-specialized models.
- **Failure Mode Analysis**: We identify recurring failure patternsincluding tokenization inefficiency, option-selection bias, repetition loops, and unintended code-switchingdemonstrating that surface-level Telugu fluency does not imply robust reasoning capability.

2 Background & Linguistic Motivation

2.1 Telugu: A Dravidian Powerhouse

Telugu is a Dravidian language spoken by approximately 80 million people across the Indian states of Andhra Pradesh and Telangana, as well as a larger global population. It possesses a rich literary tradition spanning several centuries, yet has a very limited presence on the amount of text available online and in digital forms. Unlike Indo-European languages, Telugu is morphologically agglutinative: words are formed by concatenating multiple suffixes that encode tense, case, number, aspect, and politeness. As a result, semantic units that may require several words in English often appear as a single, morphologically complex word in Telugu. For example, a concept such as *"in those houses"* can be expressed as one long word composed of multiple morphemes.

This structural property leads to high *tokenization fertility*. When processed using standard subword tokenizers, a short English sentence usually ends up expanding to three to four times as many tokens in Telugu, taking up too much of a model's context window. Since most multilingual tokenizers are optimized for English and other dominant languages, Telugu words are frequently over-fragmented into low-frequency subword or character-level units. Many Telugu roots and morphemes are absent from default vocabularies, forcing inefficient decompositions that inflate token counts and drop downstream performance.

From a linguistic perspective, Telugu also exhibits complex verb agreement, honorific systems, and a non-Latin abugida script derived from Brahmi. Reasoning tasks in Telugu, especially in legal, ethical, or commonsense domains, often require very deep contextual information and world knowledge. Earlier evidence suggests that morphologically rich, agglutinative languages systematically underperform relative to fusional languages on LLM benchmarks, likely due to a combination of data scarcity and representational inefficiency. We therefore expect even SOTA LLMs, when trained with minimal Telugu-specific data, to struggle on reasoning-intensive tasks in Telugu.

2.2 The Romanization Reality

In the Indian subcontinent, it is common practice to write Indic languages using the Latin script, especially in informal digital communication. Users frequently rely on QWERTY keyboards and phonetic transliteration, resulting in widespread romanization of native languages. For Telugu, this means that many speakers write Telugu sentences phonetically using Latin characters rather than the native script.

Crucially, there is no standardized romanization scheme. A single Telugu word may admit many valid Romanized spellings, leading to high surface-form variability. As a result, models encounter a fragmented and inconsistent input distribution when processing Romanized Telugu, further complicating tokenization and lexical alignment.

Recent analyses suggest that large language models trained primarily on English may internally rely on latent Romanized representations when processing non-Latin scripts. Even when producing output in a native script, such models may transiently encode a Latin-script approximation during intermediate reasoning (Saji et al., 2025). While this phenomenon can sometimes facilitate cross-lingual transfer, it also introduces fragility: explicitly Romanized inputs often degrade performance rather than improve it.

In TeluguEval, we also include the Romanized counterparts of the translated samples to facilitate future evaluation on Romanized scripts and resoning tasks.

3 Related Works

Much recent work has introduced multilingual and Indic benchmarks, but few focus on Telugu or on complex reasoning tasks. IndicGLUE and BHASHA-style benchmarks (Devane et al., 2025) provide classification tasks (NER, sentiment, QA) in Indian languages, primarily Indo-European ones (Maheshwari et al., 2025). For example, IndicXtreme (Doddapaneni et al., 2023) (via IndicGLUE) covers 12 languages but focuses on QA and classification; IndicMMLU (KJ et al., 2025) extends the MMLU exam to 5 Indian languages (mostly Hindi-centric). These benchmarks highlight cross-lingual transfer, but do not test multi-step reasoning in Telugu. A recent IndicParam benchmark (Maheshwari et al., 2025) (11 languages) shows larger LLMs still score <50% on low-resource tasks.

There is very minimal prior work on Telugu-specific LLM evaluation. Kishore and Shaik (2024) tested ChatGPT vs Gemini on just 20 Telugu QA questions (Kishore and Shaik, 2025), finding basic grammatical competence but noting errors on nuance. More relevantly, several studies note LLM weaknesses on morphologically complex languages (Arnett and Bergen, 2025) and script-related issues(Khullar et al., 2025). Recent research has also showed consistent drops in LLM performance on native Indian scripts vs Roman script. However, no prior work has provided a Telugu benchmark for reasoning or dissected failure modes in detail. Our work fills this gap by combining multiple reasoning domains (math, science, ethics, law) into a single evaluation, and analyzing errors such as answer bias and code-switching.

Existing ethical reasoning tasks provide the data for the ethics portion (Hendrycks et al., 2021). The CaseHOLD dataset of legal questions was introduced by Zheng et al. (2021) and contains 53K+ case holdings (Liu et al., 2025). We use subsets of these translated to Telugu. Our work differs in that we test raw model performance (no fine-tuning) across scripts, and we diagnose how multilingual/instruction-tuned models break under Telugu input.

4 The TeluguEval Benchmark

4.1 Dataset Composition

TeluguEval consists of five tasks:

- **Mathematics (GSM8K)**: 210 grade-school math word problems (70 easy, 70 medium, 70 hard) from GSM8K (Cobbe et al., 2021). These require multi-step calculation.
- **Common-Sense (Winogrande):** 200 pronoun-resolution questions (2-way MCQ) (Sakaguchi et al., 2019), testing pragmatic inference.
- **Science (ARC)**: 200 science questions (100 easy, 100 hard) from the AI2 ARC dataset (4-way MCQ), covering general science facts (Clark et al., 2018).
- **Law (CaseHOLD)**: 139 legal holding questions (MCQ) from the CaseHOLD benchmark (Zheng et al., 2021), requiring identification of legal principle.
- **Ethics (Hendrycks-Ethics)**: 200 questions (50 each from Commonsense, Deontology, Justice, Utilitarianism) from the ETHICS dataset , presenting moral dilemmas with 2-way choices.

Each English question was translated into Telugu script by bilingual experts. We also created a Romanized Telugu version by phonetic transliteration. Telugu translations preserve meaning and include culturally appropriate phrasing. For GSM8K, Google Translate was used as a rough draft and then manually corrected, ensuring exact logical consistency. Our final benchmark has 949 unique problems.

Quality Control and Agreement. Each example was translated by a single bilingual annotator and independently verified by three additional native Telugu speakers. Samples were retained only if all verifiers unanimously agreed on semantic fidelity and linguistic naturalness; any disputed items were discarded. Because disagreement cases were removed rather than adjudicated, we do not report inter-annotator agreement scores.

4.2 Translation Pipeline

Human translation was crucial because automated MT often fails on law related and ethical nuance. For example, Telugu translators noted Google often collapses fine distinctions. All questions and answer choices were verified by multiple native speakers. We made sure numeric answers remain comparable, and that Telugu options use appropriate formatting per prompt. A separate reference answer in Telugu script was recorded for

each question. For detailed information regarding the specific instructions provided to our translators and verifiers, please refer to the Appendix.

5 Experimental Setup

5.1 Models Evaluated

We evaluate eight LLMs spanning English-centric and Telugu-centric systems:

- **Global / English Models**
  - **Llama-3.1-8B-Instruct (Meta):** An 8B instruction-tuned English model with additional multilingual dialogue fine-tuning.
  - **Llama-2-7B-chat-hf (Meta):** A 7B multi-turn conversational model primarily trained for English.
  - **Qwen-3-8B (Alibaba):** An 8B model trained on large-scale English and Chinese data, with strong reasoning capabilities.
  - **Gemma-7B (English) (Google):** A 7B instruction-tuned multilingual model based on the Gemini architecture.
  - **Gemini-2.0-Flash (Google):** A large proprietary multilingual model accessed via a closed API.
- **Regional / Telugu Models**
  - **Telugu-Llama2-7B (Telugu-LLM-Labs):** A 7B Llama-2based model instruction-tuned on Telugu-specific data.
  - **Indic-Gemma-7B:** A 7B variant of Gemma-7B fine-tuned on multiple Indic languages, with a focus on Telugu.
  - **Sarvam-m-24B:** A 24B Mistral-based model with a specialized pipeline that performs intermediate reasoning in English and produces final answers in Telugu, designed for Indian languages.

All models were run via HuggingFace APIs or official interfaces. We used identical prompts (translated into Telugu where appropriate) and identical chain-of-thought prompting where allowed. For MCQs, models were instructed to pick option a, b,... or 1,2,... per task. We recorded each model's output and compared to the ground truth. For the comprehensive set of system instructions and prompt templates used across both English and Telugu evaluation modalities, please refer to the Appendix.

5.2 Evaluation Modalities

For each dataset, we ran two modes: English source (original English question), Telugu script (translated Telugu question). This isolates the effect of language on performance. Since some models have output token limits, all models were given ample context allowance; failures were typically linguistic, not truncation. We report accuracy (percentage of correctly answered questions). For multi-step math, we allow exact numeric answers and full reasoning when needed.

We also note any qualitative issues (no answer given, repetition, or answer in the wrong format) and score conservatively. Sarvam-m and Gemini sometimes output correct reasoning but in the wrong script (e.g. Telugu reasoning then English numeral), which we manually adjusted for a corrected Telugu score, shown in Results.

6 Results and Failure Analysis

6.1 Mathematics (GSM8K)

Performance on the English version of GSM8K was generally high across models. Llama-3.1-8B achieved 90.9% accuracy, Qwen-8B reached 90.5%, Sarvam-m also scored 90.5%, and Gemini-2.0 reached 96.2%. In contrast, performance dropped sharply when the same problems were presented in Telugu. Llama-3.1-8B fell to 48.1%, and Qwen-8B dropped to 75.2%. The Telugu focused models performed better than generic English models. Sarvam-m reached 75.2% accuracy, which increased to 82.4% after manually correcting answer formatting, while Gemini scored 84.8%, improving to 91.9% after correction. Smaller Telugu oriented models struggled severely, with Telugu Llama2 7B achieving 0% accuracy and Telugu Gemma 7B reaching only 19.5%.

Performance declined further on Romanized Telugu input. Llama-3.1-8B scored only 10% accuracy, corresponding to 6 correct answers out of 210. Qwen-8B achieved higher accuracy but often mixed English reasoning into its outputs, while Indic Gemma performed poorly at around 4%. We observed frequent generation failures in this setting. For example, Llama-3.1-8B sometimes re-

Model	GSM8k (En)	GSM8k (Te)	Wino (En)	Wino (Te)	ARC (En)	ARC (Te)	CaseHOLD (En)	CaseHOLD (Te)	Ethics (En)	Ethics (Te)
Llama-3 8B	90.95	48.10	66.0	56.0	91.5	36.0	58.99	33.09	59.0	53.5
Telugu Llama-2 7B	17.14	0.00	54.0	48.0	63.5	22.5	39.57	10.79	40.5	39.5
English Llama-2 7B	20.00	0.00	55.5	48.0	72.0	27.5	34.53	18.71	63.5	39.0
English Qwen-3 8B	90.48	34.29	84.0	73.5	87.5	52.5	64.75	51.08	68.0	52.0
English Gemma-7B	35.24	8.10	58.5	52.5	68.5	26.0	30.94	29.50	40.5	34.0
Telugu Gemma-7B	32.86	19.52	47.0	55.0	62.5	55.5	19.42	22.30	62.5	59.5
Sarvam-m 24B	90.48	86.67	79.5	65.0	84.0	86.5	51.80	53.96	72.5	74.0
Gemini 2.0 Flash	96.19	91.90	84.5	78.0	99.0	96.0	75.54	66.91	72.0	67.5

Table 1: Accuracy (%) of evaluated models on English (En) and Telugu (Te) versions of reasoning benchmarks.

peated the Telugu question text verbatim and never produced a numerical answer. This suggests difficulty in mapping phonetic Romanized input to the correct numerical reasoning space. Overall, the script gap is clear, with a drop of around 40 points from English to Telugu and a further drop of nearly 38 points from Telugu to Romanized Telugu.

Sarvam-m behaved differently due to its hybrid design, where reasoning is performed in English and the final answer is produced in Telugu. Its accuracy on Romanized Telugu remained close to its Telugu script performance, suggesting that the model internally converts Telugu inputs into English representations during reasoning. Across all GSM8K results, only the largest multilingual or Indic focused models, namely Sarvam-m and Gemini, maintained accuracy above 75% on Telugu inputs. All other models, including Meta's 8B models, fell below 50%.

6.2 Common Sense (Winogrande)

On the English version of Winogrande, models achieved moderate to strong accuracy. Llama-3.1-8B reached 66%, Llama-2-7B 55%, Qwen-8B 84%, Gemma-7B 58.5%, Sarvam-m 79.5%, and Gemini 84.5%. When evaluated on Telugu script, accuracy dropped to a range between 50% and 78%. For example, Llama-3.1-8B achieved 56%, Llama-2-7B 48%, Qwen-8B 73.5%, Gemma-7B 52.5%, Sarvam-m 64.5%, and Gemini 78.0%. Most models showed a drop of 5 to 15 points when moving from English to Telugu.

A prominent failure mode in this task was option bias. Several models defaulted to the same answer choice for nearly all Telugu questions. Llama-3.1-8B predicted the second option in 199 out of 200 cases. Telugu Llama2 7B consistently answered with the first option or left the answer blank, resulting in performance close to random chance. These behaviors suggest that when comprehension fails, models fall back to highly probable tokens rather than reasoning through the question. Similar positional biases have been observed in other low resource settings. Code switching was also common. Qwen-8B often appeared to translate the Telugu prompt into English internally, reason over it, and then respond, sometimes mixing languages in the final output.

When Telugu understanding failed completely, models occasionally produced unrelated or incoherent outputs. Sarvam-m and Gemini again achieved higher accuracy than other models, outperforming them by roughly 8 to 10 points, indicating more robust handling of Telugu commonsense reasoning.

6.3 Science (ARC)

On the English ARC Science questions, performance was strong for most models. Llama-3.1-8B reached 91.5%, Qwen-8B 87.5%, Gemma-7B 68.5%, Sarvam-m 84%, and Gemini 99%. Telugu inputs caused large drops in accuracy. Llama-3.1-8B fell to 36.0%, Llama-2-7B to 27.5%, Qwen-8B to 52.5%, and Gemma-7B to 26.0%. Telugu Gemma 7B performed best among the 7B models at 55.5%. Sarvam-m and Gemini remained strong, achieving 87.0% and 96.0% respectively, which is close to their English performance.

Option defaulting was again widespread. Llama-2-7B selected the same answer choice for almost all Telugu questions, indicating a breakdown in task understanding. Gemma-7B occasionally responded in a mixture of English, Telugu, and other languages. Some models produced long explanations but failed to output a final answer choice, particularly Llama-3.1-8B on Telugu inputs. In contrast, Gemini produced answers almost entirely in Telugu and was largely correct, suggesting better handling of both language and scientific knowledge.

The script gap in this task was substantial. Llama-3.1-8B dropped by roughly 55 points from

English to Telugu, while Gemma-7B dropped by about 49 points. Sarvam-m again showed much smaller degradation, with only a minor drop after correcting answer formats.

6.4 Law (CaseHOLD)

The CaseHOLD legal reasoning task proved difficult even in English. Accuracy ranged from around 30% to 64%, with Qwen-8B achieving 64.8%, Llama-3.1-8B 59.0%, Llama-2-7B 34.5%, and Gemini 75.5%. Telugu performance was substantially lower across all models. Qwen-8B dropped to 51.1%, Llama-3.1-8B to 33.1%, Llama-2-7B to 18.7%, and Gemini to 66.9%. Telugu Gemma 7B and Telugu Llama2 7B barely exceeded 20%. Sarvam-m achieved 48.2%, which increased to 53.96% after correcting answer extraction.

Failures in this domain were often severe. The smaller Telugu models frequently produced nonsensical outputs or returned zero as an answer. In many cases, models generated partial reasoning and then stopped without providing a final decision. Some runs resulted in the model looping on the question text itself. These behaviors suggest that the models lost track of the task when faced with complex legal language in Telugu. Even Sarvam-m occasionally switched back to English mid-answer despite receiving a Telugu prompt, indicating fragile alignment. Overall, the results suggest that legal reasoning in Telugu remains far beyond the reach of current models without targeted domain specific training.

6.5 Ethics (Hendrycks)

The ethics dataset consists of moral dilemmas with two answer choices. English accuracy varied by category, with commonsense tasks generally around 80 to 82%, deontology around 54 to 60%, justice around 54 to 62%, and utilitarianism between 48 and 72%. English Llama models averaged roughly 59 to 63% overall. On Telugu prompts, accuracy generally ranged from 40 to 68%. Llama-3.1-8B dropped to 53.5%, Llama-2-7B to 39.0%, Qwen-8B to 52%.

7 Discussion

Our results reveal an 8B vs. 24B paradox: smaller English models (8B) outperformably on English tasks, but a larger regionally-tuned 24B (Sarvam-m) surpasses them on Telugu tasks. Key factors include:

- **Tokenization Inefficiency**: Llama-3's tokenizer was built for English, treating Telugu as rare. As shown by (Karthika et al., 2025), applying English subwords to agglutinative scripts causes over-segmentation (high fertility). In contrast, Sarvam-m likely uses a multilingual tokenizer or larger vocab that handles Telugu subwords better. The token-overhead means Llama-3 sees fewer effective tokens of content for Telugu input, hurting reasoning.

- **Cross-Lingual Architecture**: Sarvam-m's dual-pass (think in English, answer in Telugu) approach effectively bypasses script barriers. It aligns with the idea of a latent romanization: the model may translate Telugu internally to English, reason, then transliterate back. This mitigates script gaps (Saji et al., 2025). Similarly, Gemini's top-tier training includes vast multilingual corpora, giving it robust cross-script transfer.

- **Cultural Grounding**: Some tasks (Law, Ethics) require Indian-context knowledge. English-centric LLMs lack Indian legal and moral context. Sarvam-m and Gemini, trained with Indian data, implicitly know Telugu legal terms and social norms. For example, Gemini's Telugu answers were consistently culturally appropriate (entirely in Telugu) even on complex ethics questions. This suggests semantic alignment issues: mere translation of an English model cannot capture localized knowledge.

- **Pretraining Data Volume:** (Arnett and Bergen, 2025) argue that dataset size (byte-premium adjusted) drives performance gaps. Globally pre-trained 8B models have far less Telugu content in their pretraining than Sarvam-m (24B) or fine-tuned regional LMs. With more Telugu text, Sarvam-m essentially sees more of the language during training, narrowing the gap.

In summary, our findings echo calls for tokenizer reform and regional pretraining in multilingual LLMs. English-model fine-tuning alone fails for Telugu multi-hop tasks. Future models should embed script-aware tokenizers (e.g. Indic cluster vocabularies) and scale training on Telugu corpora to avoid the fertility tax of agglutinative scripts.

8 Conclusion and Ethics Statement

8.1 Conclusion

TeluguEval shows that producing fluent Telugu text does not necessarily mean that a model can reason well in Telugu. Across mathematics, logic, and ethics tasks, large language models show large drops in accuracy when evaluated on Telugu inputs. This indicates that gains made on English benchmarks do not reliably carry over to Telugu. We observe consistent script related gaps and repeated failure patterns, suggesting that strong performance in English can hide significant weaknesses when models are used for reasoning in other languages. In contrast, only models that are explicitly designed with Telugu or Indic languages in mind, such as those with larger model sizes, cross-lingual reasoning pipelines, or region specific training data, are able to maintain high performance.

By releasing TeluguEval, the first multi-domain reasoning benchmark for Telugu, along with a detailed analysis, we provide a practical diagnostic tool for the research community. Our results suggest that languages like Telugu should not be treated as simple extensions of English centric systems. Instead, they require models that are designed to reason natively in the target language.

8.2 Implications

Our findings point to several important directions for future work:

- **Token Efficient Architectures:** Language models should incorporate script aware normalization and tokenization methods that are better suited to Indic languages, in order to reduce token inefficiency and context waste.

- **Regional Pretraining:** There is a clear need for greater investment in Telugu language data, including web text, knowledge sources, and instructional content, and for integrating this data into multilingual pretraining pipelines.

- **Sovereign Benchmarking:** Many Indian languages still lack native reasoning benchmarks. TeluguEval highlights the importance of building evaluations directly in the target language, rather than relying on translated benchmarks as proxies.

8.3 Ethics Statement

Our dataset is constructed from publicly available problems and morally neutral scenarios. All real names and personal identifiers were removed or anonymized. Model outputs were evaluated solely for correctness and reasoning behavior, and we did not observe harmful content generation within these constrained tasks. Notably, some models exhibited ungrounded refusals to benign prompts, reflecting alignment issues rather than safety concerns.

We adhere to the licenses of all underlying datasets and base models used in this work. By exposing systematic failures in Telugu reasoning, we aim to discourage the unchecked deployment of large language models in sensitive Telugu-speaking contexts without adequate evaluation and safeguards.

8.4 Limitations

TeluguEval is limited in scale, comprising 949 examples across a fixed set of academic reasoning domains. While carefully curated, it does not cover the full diversity of real-world Telugu usage. In particular, the legal reasoning subset is small by legal NLP standards. Our evaluation focuses on reasoning accuracy rather than broader measures of linguistic fluency or stylistic variation. Additionally, our translations prioritize semantic fidelity to English sources; naturally occurring Telugu queries may differ in structure and discourse style.

Despite these limitations, TeluguEval represents the first benchmark of its kind for Telugu reasoning tasks. We hope it serves as a strong foundation for future work, with the understanding that larger, more diverse, and more naturalistic Telugu datasets will be necessary to drive the next phase of progress.

Acknowledgments

The authors would like to thank the team of bilingual experts and independent verifiers for their meticulous work in translating and validating the TeluguEval dataset. Their linguistic insights were vital in ensuring the semantic and cultural integrity of the benchmark. We also express our gratitude to the Language Technologies Research Center (LTRC) at IIIT Hyderabad for providing the necessary computational resources and environment for this study.

Additionally, we acknowledge the use of AI-based tools, including ChatGPT and Grammarly, for grammatical corrections and syntactic improvements during the drafting of this manuscript.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, Suraj Racha, Nitish Kamal Singh, Ajay Nagpal, Piyush Sawarkar, Kundeshwar Vijayrao Pundalik, Rohit Saluja, and Ganesh Ramakrishnan. 2025. BhashaBench V1: A Comprehensive Benchmark for the Quadrant of Indic Domains. *Preprint*, arXiv:2510.25409.

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. *Preprint*, arXiv:2212.05409.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing Hindi Instruction-tuned LLM. *Preprint*, arXiv:2401.15006.

Revanth Gundam, Abhinav Marri, Advaith Malladi, and Radhika Mamidi. 2025a. Zero at SemEval-2025 Task 2: Entity-Aware Machine Translation: Fine-Tuning NLLB for Improved Named Entity Translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1187–1191.

Revanth Gundam, Abhinav Marri, and Radhika Mamidi. 2025b. Zero at SemEval-2025 Task 11: Multilingual Emotion Classification with BERT Variants: A Comparative Study. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1181–1186.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

N J Karthika, Maharaj Brahma, Rohit Saluja, Ganesh Ramakrishnan, and Maunendra Sankar Desarkar. 2025. Multilingual Tokenization through the Lens of Indian Languages: Challenges and Insights. *Preprint*, arXiv:2506.17789.

Manurag Khullar, Utkarsh Desai, Poorva Malviya, Aman Dalmia, and Zheyuan Ryan Shi. 2025. Script Gap: Evaluating LLM Triage on Indian Languages in Native vs Roman Scripts in a Real World Setting. *Preprint*, arXiv:2512.10780.

Katikela Sreeharsha Kishore and Rahimanuddin Shaik. 2025. Evaluating Telugu Proficiency in Large Language Models: A Comparative Analysis of ChatGPT and Gemini. *Preprint*, arXiv:2404.19369.

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. *Preprint*, arXiv:2501.15747.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025. Revisiting Epistemic Markers in Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty? *Preprint*, arXiv:2505.24778.

Ayush Maheshwari, Kaushal Sharma, Vivek Patel, and Aditya Maheshwari. 2025. IndicParam: Benchmark to evaluate LLMs on low-resource Indic Languages. *Preprint*, arXiv:2512.00333.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. RomanLens: The Role Of Latent Romanization In Multilinguality In LLMs, url=http://dx.doi.org/10.18653/v1/2025.findings-acl.1354, doi=10.18653/v1/2025.findings-acl.1354. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 2641026429. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Preprint*, arXiv:1907.10641.

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. *Preprint*, arXiv:2404.16816.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *Preprint*, arXiv:2104.08671.

Appendix

This section details the expert translation methodology, the standardized instructions provided to our bilingual verifiers, and the comprehensive prompt templates utilized across all five evaluation domains.

A Data Curation and Quality Assurance

A.1 Expert Translation Guidelines

The translation of TeluguEval was conducted by a team of proficient bilingual experts in English and Telugu. Unlike automated pipelines that often collapse fine linguistic distinctions, our human-centric approach focused on maintaining logical and semantic equivalence across diverse reasoning domains. Translators were provided with the following primary directives:

- **Semantic Integrity:** The core reasoning task (e.g., mathematical logic or pragmatic inference) must remain identical to the English source.
- **Lexical Purity:** To rigorously test the model's native language capabilities, translators were instructed to avoid English loanwords in some cases. Even for terms commonly used in code-mixed daily conversation, translators prioritized "pure" Telugu synonyms in most cases to ensure the benchmark measures deep linguistic proficiency rather than surface-level familiarity with English terms.
- **Logical Consistency:** For the GSM8K subset, translators ensured that all numeric values and their relationships remained invariant to prevent making problems easier or harder than the original source.

A.2 Verification and Multi-Stage Review

To ensure the highest data quality, each translated sample underwent a multi-stage verification process:

- **Grammar and Syntax Audit:** Three independent verifiers per dataset reviewed the translated text for grammatical correctness, natural phrasing, and proper usage of the Telugu script.
- **Semantic Alignment:** Verifiers cross-referenced every Telugu example with its English source to confirm that the semantic meaning and required reasoning steps were fully captured without distortion.
- **Romanization Accuracy:** In addition to the native script, a phonetic Romanized (Latin) version was created for each sample to facilitate testing across different input modalities. Verifiers checked these for phonetic consistency with the native script.

A.3 Cultural Adaptation

While the logical structure was preserved, phrasing was adapted to be culturally appropriate for Telugu speakers. This included the use of proper honorific systems and formal verb agreements characteristic of the language's agglutinative structure.

B Evaluation Prompts

In this section, we provide the specific system prompts and instruction templates utilized during the model inference phase. To ensure the highest quality of reasoning and adherence to output formats, the prompts were meticulously engineered and localized for each of the five reasoning domains.

B.1 Prompt Engineering and Adaptation

The prompts used in our experiments were not generic; rather, they were specifically tweaked based on the model architecture and the specific template requirements of the instruction-tuned models (e.g., Llama-3, Qwen, and Gemini). Each prompt assigns a **Strict Role** to the model (such as "Professional Mathematician" or "Legal Expert") to elicit domain-specific knowledge and maintains strict constraints on the chain-of-thought process.

For all evaluations, the English prompt templates were applied to the original English datasets

to establish baseline performance. Conversely, for the Telugu evaluations, we utilized the localized Telugu prompt templates to ensure the models were evaluated on their native-language reasoning capabilities without linguistic interference from the instructions themselves. Each prompt enforces a consistent output format consisting of a "Reasoning" field followed by a "Conclusion" field to facilitate automated exact-match scoring.

B.2 Task-Specific Prompt Templates

The following subsections provide a visual overview of the contrast between English and Telugu instructions for each evaluation task.

B.2.1 Mathematics (GSM8K)

To ensure numerical consistency and prevent models from simply repeating the question text, the GSM8K prompts enforce a step-by-step derivation format ending in a strictly formatted final answer.

Mathematics (GSM8K)

English

STRICT ROLE: You are an expert Mathematician. Your task is to solve grade-school word problems with absolute logical precision.

INSTRUCTIONS:

1. Carefully parse the numeric values and their relationships.

2. **Reasoning Field:** Show a clear, step-by-step derivation. Do not skip logical steps.

3. **Conclusion Field:** This must contain only the final numerical result.

STRICT FORMAT: Your response must end with exactly: **"Final Answer: <number>"**. Do not include units or extra text in the final answer string.

Figure 2: GSM8K English Template

Mathematics (GSM8K)

Telugu

STRICT ROLE: మీరు ఒక నిష్ణాతుడైన గణిత శాస్త్రజ్ఞుడు. ఈ గణిత సమస్యలను అత్యంత ఖచ్చితత్వంతో పరిష్కరించడం మీ బాధ్యత.

సూచనలు :

1. సమస్యలోని సంఖ్యలను మరియు వాటి మధ్య సంబంధాలను జాగ్రత్తగా గమనించండి.

2. **"reasoning" ఫీల్డ్:** సమస్యను ఎలా పరిష్కరించారో దశల వారీగా తెలుగులో వివరించండి. ఎక్కడా వివరణను దాటవేయకండి.

3. **"conclusion" ఫీల్డ్:** ఇందులో కేవలం ముగింపు సంఖ్యను మాత్రమే తెలపండి.

ఖచ్చితమైన ఫార్మాట్ : మీ సమాధానం చివరలో తప్పనిసరిగా **"తుది సమాధానం : <number>"** అని ఉండాలి. సంఖ్య తప్ప వేరే ఇతర సమాచారం ఉండకూడదు.

Figure 3: GSM8K Telugu Template

B.2.2 Common-Sense (Winogrande)

The Winogrande prompts are designed to resolve pronoun ambiguities using pragmatic inference, requiring a single sentence of logic followed by a choice between two options.

Common-Sense (Winogrande)

English

STRICT ROLE: You are a Professional Linguist and Logic Expert.

TASK: Identify the correct word to replace the underscore (_) in the sentence based on pragmatic inference.

INSTRUCTIONS:

1. Analyze the context of the sentence to resolve the pronoun ambiguity.
2. **Reasoning:** Provide a single, concise sentence explaining the logic behind your choice.
3. **Conclusion:** Choose either Option 1 or Option 2.

STRICT FORMAT: End your response exactly with: **"Final Answer: <number>"**.

Figure 4: Winogrande English Template

Common-Sense (Winogrande)

Telugu

STRICT ROLE: మీరు ఒక భాషా మరియు తార్కిక నిపుణులు.

టాస్క్: ఇవ్వబడిన వాక్యంలో ఖాళీని (_) భర్తీ చేయడానికి సరైన ఎంపికను (1 లేదా 2) ఎంచుకోండి.

సూచనలు:

1. వాక్యం యొక్క అర్థాన్ని బట్టి ఏ ఎంపిక సరిపోతుందో తార్కికంగా ఆలోచించండి.
2. **"reasoning" ఫీల్డ్:** మీ ఎంపిక ఎందుకు సరైనదో ఒక చిన్న తెలుగు వాక్యంలో వివరించండి.
3. **"conclusion" ఫీల్డ్:** కేవలం సంఖ్యను (1 లేదా 2) మాత్రమే ఇవ్వండి.

ఖచ్చితమైన ఫార్మాట్: సమాధానం చివరలో ఖచ్చితంగా **"తుది సమాధానం: <నెంబర్>"** అని ముగించండి.

Figure 5: Winogrande Telugu Template

B.2.3 Science (ARC)

The ARC prompts instruct the model to act as a science assistant, focusing on relevant scientific facts before selecting the appropriate multiple-choice label.

Science (ARC)

English

STRICT ROLE: You are a Helpful Science Assistant.

TASK: Solve the multiple-choice science question by applying scientific principles.

INSTRUCTIONS:

1. **Reasoning:** Provide a brief chain-of-thought explanation focusing on relevant scientific facts.

2. **Conclusion:** Select the single best label (A, B, C, or D).

STRICT FORMAT: End exactly with: **"Final Answer: <label>"**.

Figure 6: ARC English Template

Science (ARC)

Telugu

STRICT ROLE: మీరు ఒక సహాయకరమైన సైన్స్ అసిస్టెంట్.

టాస్క్: సైన్స్ సూత్రాలను ఉపయోగించి కింద ఇవ్వబడిన బహుళైచ్ఛిక ప్రశ్నకు సరైన సమాధానాన్ని గుర్తించండి.

సూచనలు:

1. **"reasoning" ఫీల్డ్:** సరైన ఎంపికను చేరుకోవడానికి అవసరమైన శాస్త్రీయ వివరణను తెలుగులో క్లుప్తంగా ఇవ్వండి.

2. **"conclusion" ఫీల్డ్:** సరైన ఆప్షన్ లేబుల్ (A, B, C, లేదా D) మాత్రమే ఇవ్వండి.

ఖచ్చితమైన ఫార్మాట్: మీ సమాధానాన్ని ఖచ్చితంగా **"తుది సమాధానం: <లేబుల్>"** అని ముగించండి.

Figure 7: ARC Telugu Template

B.2.4 Law (CaseHOLD)

For legal reasoning, the prompts require a concise analysis of the legal principles involved before the model identifies the correct holding numbered 0-4.

Law (CaseHOLD)

English

STRICT ROLE: You are a Legal Expert and Judicial Analyst.

TASK: Identify the legal holding (0-4) that correctly fits the provided context.

INSTRUCTIONS:

1. Carefully read the Legal Context.
2. **Reasoning:** Provide a concise analysis (max 3 sentences) explaining the selection logic.
3. **Conclusion:** State the choice clearly.

STRICT FORMAT: You MUST end the response with the choice in brackets: **"Therefore, the correct choice is [number]"**.

Figure 8: CaseHOLD English Template

Law (CaseHOLD)

Telugu

STRICT ROLE: మీరు ఒక న్యాయ నిపుణులు మరియు జుడిషియల్ అనలిస్ట్.

టాస్క్: అందించిన న్యాయపరమైన సందర్భాన్ని (Legal Context) విశ్లేషించి, దానికి సరిపోయే సరైన హోల్డింగ్ సంఖ్యను (0-4) గుర్తించండి.

సూచనలు:

1. లీగల్ కాంటెక్స్ట్‌ను నిశితంగా చదవండి.
2. **"reasoning" ఫీల్డ్:** మీరు ఎంచుకున్న సమాధానం ఎందుకు సరైనదో గరిష్టంగా 3 వాక్యాలలో తెలుగులో వివరించండి.
3. **"conclusion" ఫీల్డ్:** మీ ఎంపికను తెలపండి.

ఖచ్చితమైన ఫార్మాట్: సమాధానాన్ని తప్పనిసరిగా బ్రాకెట్లలో **[సంఖ్య]** రూపంలో ముగించాలి.

Figure 9: CaseHOLD Telugu Template

B.2.5 Ethics (Hendrycks)

Finally, the Ethics prompts guide the model through moral philosophy frameworks, ensuring the reasoning field contains a maximum of two sentences of analysis.

Ethics (Hendrycks)

English

STRICT ROLE: Moral Philosophy Expert.

TASK: Analyze the moral dilemma provided based on ethical frameworks (Justice, Utility, Deontology, or Commonsense).

INSTRUCTIONS:

1. Evaluate the scenario's implications on justice, utility, or duty.
2. **Reasoning:** Provide a maximum of 2 sentences explaining the moral analysis.
3. **Conclusion:** State the judgment index.

STRICT FORMAT: Your final result must be the label in brackets: **[number]**.

Figure 10: Ethics English Template

Ethics (Hendrycks)

Telugu

STRICT ROLE: మీరు ఒక నైతిక నిపుణులు.

టాస్క్: కింద ఇవ్వబడిన నైతిక సందర్భాలను విశ్లేషించి, తగిన నిర్ణయాన్ని తెలపండి.

సూచనలు:

1. ఈ సందర్భం లోని నైతికతను లేదా సామాజిక విలువలను గమనించండి.
2. **"reasoning" ఫీల్డ్:** మీ నిర్ణయానికి గల కారణాన్ని తెలుగులో గరిష్టంగా 2 వాక్యాలలో తెలపండి.
3. **"conclusion" ఫీల్డ్:** మీ అంతిమ నిర్ణయాన్ని తెలపండి.

ఖచ్చితమైన ఫార్మాట్: చివరిలో ఖచ్చితంగా బ్రాకెట్లలోని సంఖ్యతో **[సంఖ్య]** ముగించండి.

Figure 11: Ethics Telugu Template