Regional TinyStories: Using Small Models to Compare Language Learning and Tokenizer Performance

Anonymous ACL submission

Abstract

The 2023 TinyStories project showed that small language models (SLMs) with under 10 million parameters can generate coherent English stories when trained on carefully curated datasets. In this work, we extend the framework to Hindi, Marathi, and Bangla by using both machine-translated and LLM-generated datasets, and training SLMs up to $\approx$ 150 million parameters. We find that SLMs can produce high-quality stories in Indian languages using far fewer parameters than large models. Additionally we offer a complementary framework by using LLM-as-judge concept for *inference score based evaluation* of tokenization strategies and linguistic attributes learning. Our analysis reveals that language-specific tokenizers outperform general-purpose ones for Indian languages. Hindi models perform the strongest overall, achieving high scores in grammar, fluency, and context, supported by lower tokenization entropy and better morphological alignment. Each language exhibits different scaling behavior—Hindi benefits from wider models, Bangla emphasizes creativity with balanced setups, and Marathi requires more capacity due to its higher morphological complexity. Neural metrics based evaluations like COMET-DA and LaBSE reinforce these observations with regards to content fidelity and semantic similarity. Synthetic datasets outperform translated ones by 15–30 %. Our results advance both the practical application of SLMs to underserved languages and the theoretical understanding of neural language development. [1]

1 Introduction

Recent research on Language Models (LMs) has predominantly focused on scaling to multi-billion parameters (Brown et al., 2020; Chowdhery et al., 2022), as larger models tend to offer improved performance despite increased computational demands (Hoffmann et al., 2022b). However, the TinyStories framework by Eldan and Li (2023a) challenges this paradigm by demonstrating that Small Language Models (SLMs) with fewer than 50M parameters can perform well when trained on smaller, carefully curated datasets. This mirrors children's language acquisition, typically exposed to fewer than 100 million words by age 13 (Gilkerson et al., 2017). TinyStories established that coherent text generation can emerge from smaller architectures, language capabilities develop hierarchically, and model width correlates with knowledge retention while depth enhances contextual understanding.

Small Language Models (SLMs) surged to the fore with the BabyLM challenge (Warstadt et al., 2023a), which frames language learning under tight data budgets. Follow-up work shows their promise: Muckatira et al. (2024) demonstrate zero-shot gains from tiny vocabularies, while recent surveys argue that edge-deployable, privacy-preserving SLMs are a sustainable alternative to datacenter LLMs (Lu et al., 2025; Nguyen et al., 2024). Hugging Face anchors this shift—*TinyLlama* (1.1 B params, 1 T tokens) matches much larger baselines through rigorous curation (Zhang et al., 2024). Synthetic corpora likewise boost Arabic models (Boughorbel et al., 2024), yet mixed synthetic + TinyStories data yields only modest gains for LTG-BERT versus GPT-Neo on the original TinyStories set (Samuel et al., 2023; Black et al., 2021), underscoring data quality's primacy. Indic efforts push further: Nemotron-Mini-Hindi-4B (Joshi et al., 2025) and the 2 B-param Sarvam-1 (Sarvam AI, 2024) beat 8–9 B baselines on IndicXTREME and TriviaQA, proving sub-5 B models can deliver state-of-the-art accuracy when paired with tailored tokenizers and synthetic Indic text. Together, these results frame ultra-compact, culturally aligned SLMs as a realistic path to sovereign AI and broad digital inclusion—yet even today's "small" 5 B models remain heavy. Extending the TinyStories paradigm to

[1] Anonymous: Code & Datasets

regional tongues simultaneously provides a means to amplify under-represented languages in NLP research, drives the SLM entry barrier down, whilst shedding light on a new perspective of using SLMs as comparative tools.

Small-model experiments can reliably forecast many basic trends—like loss reduction, improved syntactic accuracy on common constructions, and general bias trajectories—via well-fitted scaling laws, letting us extrapolate from, say, 50 M to 500 M or 1 B parameters (Kaplan et al., 2020; Hoffmann et al., 2022a; Hu et al., 2020; Tal et al., 2022). However, truly emergent abilities (e.g., sophisticated reasoning or in-context learning), irregular learning curves for hard syntactic phenomena, and qualitative shifts in bias often defy such extrapolation and demand direct mid- and large-scale evaluation (Wei et al., 2022; Lu et al., 2024; Bunzeck and Zarrieß, 2024; Tal et al., 2022). In practice, one should use small models to map smooth trends, but always validate at target scale to catch surprises that only surface beyond ∼1 B parameters; however, training on such larger-sized models remains a resource constraint for us.

While the foundational TinyStories work offers compelling evidence for English SLMs, **critical questions remain**:

- Can this paradigm extend effectively to Indian languages? Does English bias in models, (Wang et al., 2024; Warstadt et al., 2023b), with English comprising a large portion of training data, limit non-English evaluation?
- Can TinyStories framework serve as a tool to analyze complexities (how difficult it is for a model to ”learn” a language) across languages, perhaps using context tracking as proxy?
- Given that the ultimate objective of training a small language model (SLM) is to maximise inference quality, could this framework serve as a rapid and cost-effective proxy for evaluating inference time smooth trends in metrics, such as grammar and fluency, while uncovering the thresholds where emergent trends like context awareness and creativity develop, when training on a specific dataset? Can these metrics in turn facilitate direct comparisons across datasets, thereby aiding in the selection of optimal data sources for scaling these trends to larger models?

To explore this, we focus on Hindi, Marathi, and Bangla—languages with substantial speaker populations and distinct linguistic features. Despite their importance, limited comparative analysis exists on their complexities in neural language modeling.

Our contributions include:

- Adapting the TinyStories SLM paradigm to these languages, detailing effective pre-training and demonstrating high inference quality with smaller models.
- Establishing a methodology for comparing linguistic complexity using TinyStories.
- Providing a novel framework for evaluating tokenizers based on SLM inference quality. Analyzing tokenization efficiency, comparing standard approaches with specialized Indian tokenizers like Sarvam and SUTRA.
- Demonstrating synthetic dataset generation outperforms translation-based approaches for inference quality.
- Highlighting lexical differences in training data, emphasizing variety despite semantic equivalence, and showing limitations of metrics like ROUGE for morphologically rich languages.
- Releasing a total training data of 10M synthetic and translated stories in three Indian languages and SLMs weights for the community use.

These findings suggest effective language modeling for Indian languages may not require massive architectures but rather task-specific quality datasets, potentially democratizing access to language technology for underrepresented languages.

2 Methodology: Data Generation, Training and Evaluation Experiments

2.0.1 Translated Data

Previous work by Doshi et al. (2024) showed that machine-translated filtered data can train Indian language models matching the performance of those trained on native data. Similarly, NLLB-3B MT was used to translate TinyStories into Arabic (Boughorbel et al., 2024). Following this established approach, we translated the complete TinyStories dataset (2.0M stories) from English to Hindi, Marathi & Bangla using NLLB-200-3B and Google Translate. (See. Appendix D).

2.1 Training Data Preparation

Our research extends the TinyStories framework (Eldan and Li, 2023a) to explore simple, constrained narratives in multiple Indian languages by training SLMs on data generated by either: (i)

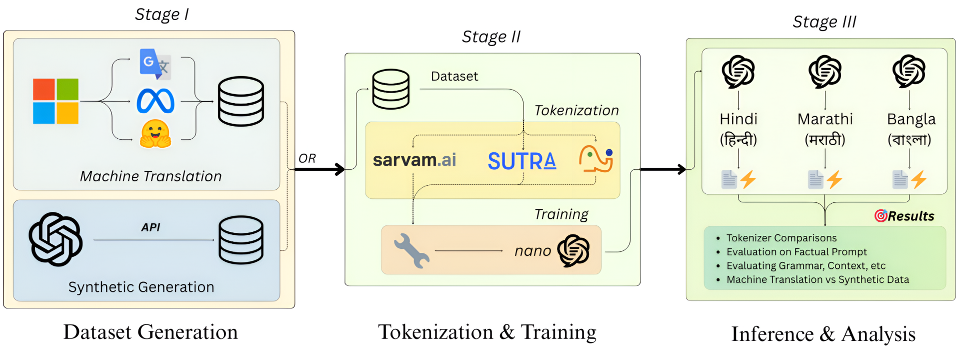

Figure 1: Regional-TinyStories Pipeline Overview

Translating the original English TinyStories dataset (Eldan and Li, 2023b) into Indian languages, or (ii) Generating synthetic data using SOTA LLMs (Fig. 1).

2.1.1 Synthetic Data

For all 3 languages, we follow the same procedure. Starting with vocabulary generation, where we curated lists of over 300 nouns, verbs, and adjectives per language (Sec. E.1. Using this vocabulary, and an optimal prompt template (Sec. E.3, E.4) we devlop a custom duplicate avoidance algorithm (Sec. E.2) to finally generate a corpus of 2M unique prompts which are inferenced by GPT-4o-mini to produce corresponding 2M stories. (Sec. E.5

2.2 Training Data Evaluation

Analysis of synthetic training data shows traditional evaluation metrics struggle with Indian languages (Appendix. F). ROUGE (Lin, 2004) scores registered zero for semantically similar Bangla stories, highlighting limitations in non-English text evaluation (Sec. F.1). BERTScore (Zhang et al., 2020) values near 1.0 confirmed strong semantic equivalence, while BLEU scores (Papineni et al., 2002) averaged just 0.078. METEOR (Banerjee and Lavie, 2005) offered a middle-ground assessment, averaging 0.153, by recognizing synonyms and variations (Sec. F.1.2, Sec F.5). This pattern—high semantic similarity with low lexical overlap—indicates our dataset features diverse expressions of similar concepts. Morphologically rich Indian languages allow extensive variation in expressing equivalent meanings. This ”anomaly” is infact a strength: Our stories maintain semantic coherence while exhibiting linguistic diversity, ideal for robust language modeling. Models learn to understand concepts through varied vocabulary instead of memorizing phrases, highlighting both the challenge of evaluating Indian language generation and the advantage of our approach, which yields semantically coherent and lexically diverse examples for improved language understanding. (Sec. F.3).

2.3 SLM: Build, Train, Inference & Evaluate

The open-source, well regarded, Indic tokenizersSarvam-1 (Sarvam, 2024) and SUTRA-mlt-256-v2 (Bendale et al., 2024) incorporate language-specific tokens, while OpenAI’s Tiktoken (OpenAI, 2024) served as baseline comparison (Tokenizer details C.4). Sarvam-1 achieves near-English token fertility rates (1.4-2.1 tokens/word) for Indic scripts, significantly improving efficiency compared to traditional multilingual LLMs.

Our SLMs are built from scratch in PyTorch using the nanoGPT codebase (Karpathy, 2022), implementing decoder-only transformers with 8 attention heads at various parameter sizes (Sec. C.1). All models trained for 5000 epochs with 2.5% data reserved for testing (Details. C.2). Following training, the SLMs were inferred (Sec . B.1) to produce 3 unique stories per 1000 prompts (non-occurring in the training data). As per the seminal work, we use the LLM-as-judge (Eldan and Li, 2023a; Zheng et al., 2023) framework to evaluate each story. (Details. B.1.2).

3 Results & Discussions

3.1 LLM-as-judge Evaluation Scores

3.1.1 Model Architecture and Scaling Dynamics

We showcase results for Hindi (Fig. 1 while the extended results for Hindi, Marathi, & Bangla models, trained using the Sarvam tokenizer on synthetic data follow similar trends and can be found at Tabs. 7, 8, 9).

We thoroughly explore the relationship between Model Architecture and Evaluation Metrics (Sec. G.5), studying divergent trends and sensitivity of each metric). We observe a strong increase across all evaluation metrics, up to an optimal configuration of 512 embedding dimensions and 6 layers (54M params), after which performance plateaus.

3.1.2 Scaling of Individual Metrics

Across Hindi, Marathi, and Bangla, we observe a *hierarchical emergence of capabilities*: grammatical correctness plateaus at the highest level, fluency improves next, whereas completeness and—most critically—context awareness lag behind. This divergence shows that evaluation metrics scale *non-uniformly* with model capacity, making context fidelity the primary bottleneck for further quality gains (Sec. G.4. We conduct various in-depth studies Correlation (Sec. G.6), Variance Distribution (Sec. G.7), Score-Consistency (Sec G.7.2, and Gap (Sec. G.8) to further uncover trends.

3.1.3 Contextual Emergence of Linguistic Capabilities

Recent evidence demonstrates that ostensibly *emergent* abilities—context tracking, discourse-level reasoning, and related competences—are attainable well below the billion-parameter scale (Muckatira et al., 2024). Our experiments corroborate this claim: models containing merely 10 M parameters already reproduce narrative flow and grammatical accuracy comparable to substantially larger systems. Basic linguistic skills therefore appear to saturate early, within the 4–10 M–parameter band, whereas *context awareness* remains stubbornly elusive until the representational capacity afforded by $\sim$ 512-dimensional embeddings (roughly 41–73 M parameters), with sharply diminishing returns beyond 1024 dimensions.

Statistical analyses and targeted manual reviews (Items. 4, 3, 1, and 3) converge on the conclusion that context fidelity constitutes the single most informative metric for holistic model assessment. Its delayed maturation underscores a *hierarchical emergence* of capabilities: foundational grammar stabilises first, fluency follows, and only then does coherent, prompt-aligned discourse solidify.

In this process, we successfully show how our SLMs can be used as a rapid and cost-effective proxy for evaluating inference time smooth trends in metrics, such as grammar and fluency, while uncovering the thresholds where emergent trends like context awareness. (Item. 1) While, SLMs as a proxy for evaluating datasets is discussed in Sec. 3.3.

This trajectory mirrors developmental patterns observed in human language acquisition, where simpler morphological and syntactic regularities precede more sophisticated narrative competencies. Such parallels lend support to usage-based theories of language learning, positing that emergent linguistic representations spring from domain-general faculties—intention reading, pattern detection, and social interaction—rather than a pre-wired, specialised “language module” (Tomasello, 2003). Consequently, future optimisation should prioritise training signals that strengthen long-range context integration, as further scaling alone yields limited dividends once grammatical and fluency plateaus are reached.

3.2 Inference Examples & Factual Performance

3.2.1 Inference Analysis

The methodology for generating and evaluating stories from models is described in Sec. 2.3 (In Brief) & Sec. B (In Detail).

Examples of the Hindi 54 M model can be found in Table. 2 (see Tables. 13, 14 & 15 for more examples), where we observe that, although short, the stories are narratively complete with a strong moral.

We crucially uncover, in Sec. B.2, how Marathi Models lag behind Hindi and Bangla models in terms of context awareness. These claims are further bolstered through our statistical analyses (Sec. 2 & Sec. G.10)

Comparison with reference SOTA LLMs provides valuable insights into the current capabilities and limitations of our approach. For softer-trends (Item. 1), we demonstrate that despite having a parameter count nearly 1 M times lower than GPT-4o, *we successfully generate coherent and fluent stories with clear messages*. Emergent, threshold-based trends, however, clearly lag behind confirming the existence of further ”jumps”/”rewards” in performance at larger model size.

3.2.2 Factual Analysis

SLM’s ability to correctly answer factual questions holds key as they probe context-tracking, logical comprehension, and hallucination propensity in ways rote recall cannot. Factual performance grows strongly with Increasing Model Size. While we observe, Hindi > Marathi > Bangla, despite Bangla Models show higher context awareness. (Sec . B.3)

3.3 Translated vs Synthetic Datasets

The 54M model with Sarvam tokenizer revealed significantly lower performance when trained on translated versus synthetic data. Models using translated data showed higher evaluation loss and poorer inference scores despite identical training duration. Hindi models scored just 6.30 overall with translated data compared to 8.16 with synthetic data. (Appendix. D.2).

This aligns with previous findings (Boughorbel et al., 2024) and can be explained based on the following : (1) Cultural biases where source language elements transfer inappropriately to target languages (Holmström et al., 2023), (2) Grammatical and stylistic challenges where translations fail to capture language-specific structures and conventions (Zhang and Toral, 2019), and (3) Translation-induced noise that impedes next-token prediction compared to original text generation (Boughorbel et al., 2024).

Appropriately comparing evaluation scores, we successfully show how our Regional TinyStories framework can be used as a proxy to compare datasets via inference time evaluation of models trained on said datasets. (Item. 1)

3.4 Regional Tokenizers Perform Better

Table. 4 compares three tokenizers—Sarvam, SUTRA, and Tiktoken—across Hindi, Marathi, and Bangla using both quantitative and qualitative metrics based on stories generated by our 54M parameter models. Our analysis reveals a striking pattern: Tiktoken consistently achieves the lowest evaluation loss across all languages (Hindi: 0.149, Marathi: 1.167, Bangla: 0.135), suggesting superior perplexity minimization. However, this advantage doesn't translate to generation quality, where Tiktoken underperforms on all subjective dimensions. Indian language-specific tokenizers demonstrate superior performance in generation quality. Sarvam achieves similar overall scores for all languages (Hindi: 8.158, Marathi: 8.296, Bangla: 8.016). Particularly excelling in context understanding and narrative completeness. SUTRA follows closely, with strengths in grammatical accuracy. It is important to consider that Marathi fares significantly worse with regard to contextual awareness as compared to Hindi and Bangla.

The performance gap is most pronounced in context awareness (+0.56 points average for Sarvam over Tiktoken) and fluency (+0.63 points average). This suggests *regionally specialized tokenizers better capture semantic cohesion, idiomatic expressions, and structural nuances*. These findings align with research showing general-purpose tokenizers introduce significant biases in non-English languages, requiring up to 15 times more tokens for equivalent content (Petrov et al., 2023).

The superior performance of language-specific tokenizers can be attributed to several factors: (1) More efficient subword segmentation aligned with morphological boundaries, (2) Better handling of script-specific features in Devanagari (Hindi/-Marathi) and Brahmi (Bangla) scripts, and (3) Vocabulary coverage optimized for the linguistic distributions of these languages. These advantages are particularly evident in the grammar scores, where both Sarvam and SUTRA demonstrate robust handling of morphological and syntactic features specific to Indian languages.

3.5 Mechanistic Evaluations of Inference Results

We use a dual-perspective approach to quantify the linguistic complexity of Hindi, Bangla, and Marathi, focusing on tokenization strategies. This is to further explain the observations of our inference-based evaluation framework.

3.5.1 Information-Theoretic Analysis

To assess tokenization quality and language complexity, we computed Rényi entropy (Zouhar et al., 2023), which measures uncertainty and diversity in token distributions. This analysis was conducted for Hindi, Bangla, and Marathi using Sarvam and SUTRA tokenizers. Rényi entropy quantifies information content in tokenized distributions, with α controlling sensitivity to rare tokens.

Our findings reported in Table. 5, indicate Marathi has the highest entropy, suggesting a complex morphological structure and greater token variability, possibly leading to lower evaluation scores. Sarvam shows lower entropy across languages, indicating efficient subword capture and compact tokenization, while SUTRA's higher entropy reflects a diverse strategy, offering richer representation but increased complexity. This may explain Sarvam's superior model performance.

Entropy was examined at α levels of 0.5, 1.0, and 2.0 (Appendix. H). At α = 0.5, which emphasizes

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
64	2	4.46	1.408	5.665	6.826	7.217	7.472	7.969	7.030
64	6	4.65	1.182	6.412	7.122	7.314	7.901	8.446	7.439
64	12	5.00	1.057	6.374	7.227	7.390	7.959	8.450	7.480
512	2	41.00	0.654	7.054	7.661	7.705	8.427	8.746	7.919
512	6	53.00	0.518	7.734	7.783	7.806	8.554	8.912	8.158
512	12	73.00	0.519	7.572	7.659	7.718	8.458	8.862	8.054
1024	2	94.00	0.581	7.344	7.798	7.829	8.516	8.825	8.062
1024	7	153.00	0.513	7.695	7.806	7.830	8.580	8.910	8.164
GPT-4o-mini	-	-	-	8.915	9.725	8.493	9.803	9.835	9.354
GPT-4	-	-	-	9.368	9.899	8.749	9.880	9.898	9.559
Gemini-1.5-flash	-	-	-	9.204	9.906	8.820	9.794	9.802	9.505

Table 1: **Hindi** - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

Model	Completion	Evaluation
Hindi		
Prompt	**Indian Regional:** सूर्योदय पर, दो सबसे अच्छे दोस्त नदी के कनिारे अजीब पैरों के नशिान का पीछा कर रहे थे। **Translation:** At sunrise, two best friends were following some strange footprints along the riverbank.	
Medium *54M* $E = 512$ $L = 6$ $H = 8$	**Indian Regional:** सूर्योदय पर, दो सबसे अच्छे दोस्त नदी के कनिारे अजीब पैरों के नशिान का पीछा कर रहे थे। जैसे ही सूरज की करिणें नदी पर पड़ीं, पानी में एक अद्भुत चमक दखिाई दी। सभी दोस्त एक-दूसरे को देखकर मुस्कुरा रहे थे। फरि अचानक, मोहन ने देखा कि पानी में कुछ रंग-बरिंगे फूल तैर रहे हैं। उसने सोचा, "अगर मैं इन फूलों को पकड़ लूं, तो यह बहुत मजेदार होगा!" उसने अपने दोस्तों को बुलाया और सबने मलिकर फूलों को पकड़ने की कोशशि की। खेल-खेल में, मोहन और उसके दोस्तों ने देखा कि नदी के कनिारे एक छोटा सा रहस्य छपिा हुआ है। जब उन्होंने उन फूलों को पकड़ा, तो उन्हें पता चला कि वे सरि्फ सुंदर थे और उनमें जादुई शक्तियाँ थीं। सभी दोस्त उस दनि को कभी नहीं भूल पाए और हर सुबह सूर्योदय के समय फूलों के साथ खेलते रहे। **English Translation:** *At sunrise, two best friends were following some strange footprints along the riverbank.* As soon as the sun's rays touched the river, a magical sparkle appeared on the water. All the friends looked at each other and smiled. Suddenly, Mohan noticed some colorful flowers floating in the water. He thought, "If I could catch these flowers, it would be so much fun!" He called his friends, and together they tried to catch the flowers. While playing, Mohan and his friends discovered that a small secret was hidden along the riverbank. When they caught the flowers, they realized that they were not just beautiful—they had magical powers. None of the friends could ever forget that day, and from then on, they played with the flowers every morning at sunrise.	***Overall: 8.6*** *Context: 8* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*

Table 2: Hindi Inference Example

rare tokens, Marathi had the highest entropy. At α = 1.0 (Shannon entropy), moderate convergence is observed, but Marathi remained higher. These patterns confirm tokenization complexity differences among the languages.

3.5.2 Morph Score

We evaluated morphological fidelity using MorphScore to assess alignment between tokenizer outputs and linguistic morphemes, following Arnett and Bergen (2024). A *morpheme* is the smallest unit of language with meaning, serving as a basic building block for words. Morphologically-annotated evaluation sets were constructed for each language.

As shown in Table 6, Sarvam scores slightly higher for Hindi and Bangla, suggesting better morphological boundary preservation. Significant variation is seen across languages, with Bangla scoring notably lower than Hindi and Marathi, indicating suboptimal tokenization for Bangla. Hindi (Devanagari script, Indo-Aryan) and Marathi (also Devanagari) share more subword roots than Bangla (Brahmic script variant), possibly giving Hindi/Marathi an advantage with a shared tokenizer. While higher MorphScores generally align with better model performance, Bangla models outperform Marathi, implying factors like Rényi entropy may play a significant role, warranting further research.

3.5.3 Cross-Linguistic Complexity Analysis

Our analyses offer insights into language complexity across Marathi, Bangla, and Hindi. Marathi exhibits higher complexity in token distribution (highest Rényi entropy), while Bangla faces challenges in morphological boundary recognition (lowest MorphScore), which perhaps is the root cause of poor factual performance (Sec. 3.2.2 & Sec. B.3). Hindi shows moderate metrics, aiding efficient tokenization and superior model performance.

Bafna and Žabokrtský (2022) noted Marathi's agglutinative nature, allowing suffix stacking, unlike Hindi's token separation and Arnett and Bergen (2024) showed higher Rényi entropy for agglutinative languages.

Although our inference evaluations show Marathi models having higher performance (Ap. A.1), uncovering the importance of context awareness

Prompt *Eng. Translation*	**5M** *Hindi*	**54M** *Hindi*	**157M** *Hindi*	**5M** *Marathi*	**54M** *Marathi*	**5M** *Bangla*	**54M** *Bangla*
When Ram came back home, he was very tired, so he	was very happy. बहुत खुश था।	wanted to relax. आराम करना चाहता था।	he went to his bed to sleep. सो जाने के लिए अपनी बसि्तर पर चला गया।	slept. झोपला.	went to sleep. झोपायला गेला.	sat down. বসে পড়ল।	went to sleep. সে ঘুমাতে গলে।।
Jack and Lily saw a rainbow after a rainy day. They were amazed by the colors. Jack said, "Look, Lily. A rainbow has	a lot of beauty! कितनी सुंदरता है!	seven colours! सात रंग हैं!	seven colours! सात रंग हैं!	a lot of beauty! किती सुंदरता आहे!	many colours किती रंग आहेत!	a lot of beauty! কত সুন্দর!	beautiful colours! একটি রং আছে!

Table 3: Cross-lingual performance across key model sizes on **factual prompts** — *GPT-4 and better = 6/6*
Evaluation Scheme: Red = 1/6, Light Red = 2/6, Light Yellow = 3/6, Yellow = 4/6, Light Green = 5/6 & Green = 6/6
Left to Right, Red to Green, Arbitrary to Factual Contextual Completion

Tokenizer Name	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi							
Sarvam-1	0.518	7.734	7.783	7.806	8.554	8.912	8.158
SUTRA-mlt-256-v2	0.522	7.548	7.449	7.584	8.292	8.875	7.950
Tikoken	0.149	6.974	7.106	7.360	7.889	8.681	7.602
Marathi							
Sarvam	1.662	7.154	8.127	7.902	8.854	9.146	8.296
SUTRA	1.824	7.223	7.862	7.633	8.602	9.024	8.069
Tiktoken	1.167	6.514	7.442	7.437	8.105	8.851	7.670
Bangla							
Sarvam	0.569	7.507	7.645	7.693	8.420	8.816	8.016
SUTRA	0.608	7.614	7.374	7.595	8.212	8.845	7.928
Tiktoken	0.135	7.118	6.989	7.358	7.778	8.614	7.572

Table 4: **Tokenizer performance** across Hindi-Marathi-Bengali
Sarvam Tokenizer — 54M (E=512, L=6) Model — Lighter is better

Tokenizer	Hindi	Bangla	Marathi
Sarvam	6.2852	6.3579	6.5449
SUTRA	7.1530	7.4135	7.7620

Table 5: Rényi entropy ($\alpha = 2.5$) for Hindi, Bangla, and Marathi using Sarvam and SUTRA tokenizers.

Language	SUTRA	Sarvam
Hindi	0.7268	0.7276
Bangla	0.3002	0.3194
Marathi	0.6671	0.6620

Table 6: MorphScore evaluation results comparing SUTRA and Sarvam tokenizers across three Indic languages. Higher scores indicate better alignment with morphological boundaries.

(Sec. 3.1.3 & Sec. G.10) and said poor performance of Marathi models (Sec. 3.1.3, Items. 4 & . 2), along with the least Rényi entropy indicate that SLMs face difficulty in "learning" a Marathi corpus, *validating our hypothesis of using SLM inference scores to assess language complexity*. (Item.1)

We can now conclude that *equally weighted averaged* evaluation metrics do not paint the complete picture, and complexity, when defined as the ability of a model to "learn" a language, shows the trend of *Hindi* $>$ *Bangla* $>$ *Marathi*. (similar. G.10)

Prior work (Arnett and Bergen, 2024) suggests that language difficulty for models isn't solely due to morphology. While fusional languages often outperform agglutinative ones, the performance gaps could be driven by differences in dataset size and encoding efficiency—closely tied to the predictability of token sequences, which we show can be quantified using measures like Rényi entropy—rather than only through tokenizer alignment or quality.

4 Additional Benchmarking (Ap. I)

In the absence of human evaluation, we provide supplementary assessments to evaluate the semantic quality of SLM vs GPT-4o stories by employing a combination of advanced neural metrics. COMET-DA (Rei et al., 2022) measures *content fidelity* (match between the semantic structure of SLM vs GPT-4o stories) while LaBSE (Language-Agnostic BERT Sentence Embedding) (Feng et al., 2022) measures language-agnostic *semantic similarity* between stories, but neither captures narrative coherence or creativity.

This embedding-based approach allows another evaluation even when lexical variations or paraphrasing are present. Across all three languages, model performance generally improves with increased parameter count, *mirroring observations in*(Appendix. A), showing steady gains up to medium-sized models (54-95M parameters), after which performance plateaus or slightly decreases at the largest checkpoints.

COMET-DA scores climb from 0.56 to 0.66 (Hindi) and 0.68 to 0.76 (Bangla) before saturating, while Marathi tops out at 0.59, echoing its lower LLM judged scores. LaBSE shows that all three languages already achieve strong topical alignment by 40-60M parameters (0.75–0.78), implying that the residual gap in COMET-DA and in Tables. 7, 9, 8 are driven by surface fluency and discourse, not missing content. COMET-DA rises sharply with model size up to ~50M parameters, indicating better content fidelity, then shows diminishing returns, while LaBSE reveals that strong semantic overlap (0.75–0.78) is attained even by smaller models, with only minor language-specific differences thereafter.

In sum, these metrics support the LLM as judge findings and reinforce our conclusion that **tokenizer quality and data curation**, rather than further scale beyond ~50M parameters, are the **key drivers of story quality** in Indic SLMs.

5 Related Work: Multilingual Language Modeling

Our findings align with the "curse of multilinguality" identified by Chang et al. (2024), considering how our focused Hindi, Bangla, and Marathi SLMs achieve comparable performance to significantly larger multilingual models.

Chang et al. (2024) trained multiple $\leq$ 45M parameter, GPT2–style SLMs across 252 languages and showed that moderate multilingual mixing benefits closely related low-resource languages ($\approx\leq$ 10M tokens) but progressively hurts high-resource languages, confirming that small-capacity models cannot accommodate large multilingual corpora without sacrificing performance.

Alternatively, solutions like X-MOD (Pfeiffer et al., 2022) and X-ELM (Blevins et al., 2024) demonstrate that linguistic similarity enables effective cross-lingual transfer through specialized modules. X-MOD's modularity is relevant to our findings on different language-specific learning patterns: Hindi excels in context/grammar, Bangla emphasizes creativity, and Marathi requires larger models.

Our Indic tokenizer comparison supports specialized approaches, with Sarvam and SUTRA outperforming general-purpose alternatives. Regional TinyStories could integrate modular components for language families (e.g., Indo-Aryan, Dravidian), creating efficient models that respect linguistic typology while reducing computational demands, particularly valuable for low-resource scenarios.

6 Conclusion

We extend the TinyStories paradigm to Indian languages, showing that SLMs with 5–54M parameters can generate fluent, coherent stories in Hindi, Bangla, and Marathi. Our best-performing 157M model achieves over 90% of GPT-4o's performance on key linguistic metrics, despite being orders of magnitude smaller.

Our results show that effective modeling of low-resource, morphologically rich languages does not require massive architectures. Instead, targeted data curation and high-quality, language-specific tokenizers (e.g., Sarvam, SUTRA) are more decisive. Hindi models benefit from wider embeddings, Bangla emphasizes creativity with balanced configurations, and to compensate for poor context awareness, Marathi needs more parameters due to its agglutinative nature and token entropy.

We introduce an inference-based evaluation framework that benchmarks performance and reveals language-specific learning patterns and complexity. This approach, validated by LLM-based evaluation and neural metrics (COMET-DA, LaBSE), offers a scalable method to study under-represented languages. Our findings support a data- and tokenizer-first approach in multilingual NLP, enabling access to language technology for under-served regions.

Crucially, our findings establish that SLMs serve as cost-effective, efficient proxies for dataset and tokenizer benchmarking, while simultaneously enabling analyses of linguistic complexity—thereby opening a promising research trajectory for Regional Small Language Models.

Limitations

- **No human evaluations:** We rely on LLM-as-judge frameworks (e.g., GPT-4o), which remain imperfect proxies for human judgment (Chen et al., 2024), especially in morphologically complex and culturally diverse languages. Even limited human annotations could strengthen evaluation reliability. Nevertheless we would like to emphasize the sample size of 3000 inferences per model, which provides stable average scores even if variance across runs is unmeasured. Multiple evaluation metrics (LLM-as-judge, COMET-DA, LaBSE), also show converging trends across language and tokenizer configurations.

- **Single model runs:** Due to compute constraints, each configuration was trained and evaluated once. This limits our ability to report statistical variance or confidence intervals, which should be addressed in future work.

- **Evaluation-model overlap:** GPT-4 variants were used for both generating synthetic data and judging inference quality, which may introduce evaluator bias. Future work should use independent LLMs or human raters to mitigate this concern.

- **Limited tokenizer baselines:** While Sarvam and SUTRA outperform general-purpose tokenizers like Tiktoken, other strong Indic tokenizers (e.g., IndicBERT) were not tested and may offer further insights.

- **Metric limitations:** COMET-DA and LaBSE capture semantic similarity but fail to evaluate narrative coherence or creativity. Incorporating story-aware or structure-aware evaluation metrics is a promising direction.

- **Architectural scope:** All models are decoder-only transformers. Exploring hybrid or sparse architectures (e.g., Mixture-of-Experts) may yield better performance under compute constraints.

- **No multilingual SLM training:** We train separate models for each language. Multilingual training across related languages (e.g., Hindi-Marathi) may unlock shared learning benefits and improve data efficiency (Chang et al., 2024).

We remain committed to addressing these limitations, with ongoing efforts particularly focused on developing a human evaluation study that rigorously accounts for ethical and privacy considerations.

Impact Statement

Opportunities and Challenges

Our work on generating children's stories in Indian regional languages offers key opportunities for educational access and cultural preservation. Open-weight models deployable across environments (from edge devices to the cloud) can help address the scarcity of literature in underserved languages.

This technology can support early childhood literacy, especially in rural areas with limited publishing infrastructure. However, successful deployment requires attention to cultural nuances, content moderation, and preservation of regional storytelling traditions.

Recent multilingual benchmarks highlight a practical blueprint for low-resource language modeling: starting with a TinyStories-style corpus (LLM or human-generated) and using language-specific tokenization and training. Our approach demonstrates this for Hindi, Marathi, and Bangla, and similar results have been shown for African languages like Swahili and Yoruba (LelapaAI, 2024).

To ensure responsible use, we recommend:

- Robust review mechanisms involving language experts
- Clear guidelines for cultural appropriateness
- Metrics to measure educational impact

This technology should complement and not replace traditional storytelling, working alongside educators and cultural experts to preserve authenticity while leveraging the benefits of AI-generated content.

Code and model repository

We publicly release all our codes and datasets used for the experiments described in this work, as well as all trained models, to allow for external validation and reproducibility:

1. Anonymous Codebase: GitHub: Regional-TinyStories
2. Anonymous Datasets: HF: Regional TinyStories

References

Catherine Arnett and Benjamin K. Bergen. 2024. Why do language models perform worse for morphologically complex languages? *Preprint*, arXiv:2411.14198.

Niyati Bafna and Zdeněk Žabokrtský. 2022. Subword-based cross-lingual transfer of embeddings from Hindi to Marathi and Nepali. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 61–71, Seattle, Washington. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Abhijit Bendale, Michael Sapienza, Steven Ripplinger, Simon Gibbs, Jaewon Lee, and Pranav Mistry. 2024. Sutra: Scalable multilingual language model architecture. *Preprint*, arXiv:2405.06694.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

Sabri Boughorbel, MD Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. *Preprint*, arXiv:2405.14277.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shades of blimp: Syntactic learning curves. In *Proceedings of the CLASP Workshop at EACL*.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Ljiljana Dolamic and Jacques Savoy. 2010. Comparative study of indexing and search strategies for the hindi, marathi, and bengali languages. *ACM Transactions on Asian Language Information Processing*, 9(3).

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023a. roneneldan/tinystories. `https://huggingface.co/datasets/roneneldan/TinyStories`.

Ronen Eldan and Yuanzhi Li. 2023b. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judy K. Montgomery, Charles R. Greenwood, D Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26 2:248–265.

Jordan Hoffmann, Sebastian Borgeaud, Marcin Bosma, and et al. 2022a. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022b. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Jennifer Hu, Xiaodong Lu, Julie Dillon, Tal Linzen, Evelina Fedorenko, and Chris Dyer. 2020. Systematic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3398–3410.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *Preprint*, arXiv:2410.14815.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andrej Karpathy. 2022. nanogpt.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

LelapaAI. 2024. InkubaLM: A Small Language Model for Low-Resource African Languages. Blog post.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Sheng Lu, Biswajit Majumder, Seungwhan Min, and et al. 2024. Are emergent abilities truly emergent? In *ACL Findings*, pages 999–1010.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. 2025. Small language models: Survey, measurements, and insights. *Preprint*, arXiv:2409.15790.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent abilities in reduced-scale generative language models. *Preprint*, arXiv:2404.02204.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, and 9 others. 2024. A survey of small language models. *Preprint*, arXiv:2410.20011.

OpenAI. 2024. Tiktoken.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Preprint*, arXiv:2305.15425.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Sarvam. 2024. Sarvam 1 : The first indian language llm.

Sarvam AI. 2024. Sarvam-1: The first indian language small language model. `https://www.sarvam.ai/blogs/sarvam-1`. Accessed 20 May 2025.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yarden Tal, Amir Zeldes, and Yoav Goldberg. 2022. Fewer errors, but more stereotypes? bias trends with model scale. In *Proceedings of the Workshop on Gender Bias in NLP*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Michael Tomasello. 2003. *Toward a Psychology of Language Acquisition*, pages 323–328. Harvard University Press.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *Preprint*, arXiv:2411.03350.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2301.11796.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023b. Blimp: The benchmark of linguistic minimal pairs for english. *Preprint*, arXiv:1912.00582.

Jason Wei, Yi Tay, Rishi Bommasani, and et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Shiyue Zheng, Tuhin Ghosal, Ethan Chi, and et al. 2023. Judging llm-as-a-judge: Comparison and calibration of llms for evaluating open-ended text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

Appendices

Appendix Index

A Complete Hyperparameter and Tokenizer Comparison Tables

A.1 Complete Hindi, Marathi & Bangla Hyperparameter Comparisons

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
64	2	4.46	1.408	5.665	6.826	7.217	7.472	7.969	7.030
64	6	4.65	1.182	6.412	7.122	7.314	7.901	8.446	7.439
64	12	5.00	1.057	6.374	7.227	7.390	7.959	8.450	7.480
128	2	9.00	1.022	6.473	7.251	7.501	7.978	8.464	7.533
128	6	10.00	0.864	7.008	7.401	7.559	8.222	8.699	7.778
256	2	19.00	0.819	6.861	7.566	7.665	8.303	8.659	7.811
256	6	27.00	0.618	7.257	7.645	7.653	8.433	8.797	7.957
512	2	41.00	0.654	7.054	7.661	7.705	8.427	8.746	7.919
512	6	53.00	0.518	7.734	7.783	7.806	8.554	8.912	8.158
512	12	73.00	0.519	7.572	7.659	7.718	8.458	8.862	8.054
768	2	66.00	0.605	7.369	7.664	7.685	8.441	8.822	7.996
768	6	95.00	0.517	7.711	7.773	7.820	8.551	8.906	8.152
1024	2	94.00	0.581	7.344	7.798	7.829	8.516	8.825	8.062
1024	7	153.00	0.513	7.695	7.806	7.830	8.580	8.910	8.164
GPT-4o-mini	-	-	-	8.915	9.725	8.493	9.803	9.835	9.354
GPT-4	-	-	-	9.368	9.899	8.749	9.880	9.898	9.559
Gemini-1.5-flash	-	-	-	9.204	9.906	8.820	9.794	9.802	9.505

Table 7: **Hindi** - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
64	2	4.46	2.680	6.046	7.119	7.292	7.607	8.332	7.279
64	6	4.65	2.343	6.850	7.430	7.389	8.036	8.809	7.703
64	12	5.00	2.224	6.830	7.520	7.465	8.094	8.813	7.744
512	2	41.00	1.773	6.474	8.005	7.801	8.726	8.980	7.997
512	6	54.00	1.662	7.154	8.127	7.902	8.854	9.146	8.236
512	12	73.00	1.654	6.992	8.003	7.814	8.757	9.096	8.132
1024	2	94.00	1.682	7.290	7.689	7.199	8.193	8.855	7.845
1024	7	157.00	1.640	7.641	7.697	7.200	8.257	8.940	7.947
GPT-4o-mini	-	-	-	8.630	9.459	8.416	9.359	9.441	9.061
GPT-4	-	-	-	9.329	9.812	8.770	9.728	9.749	9.477
Gemini-1.5-flash	-	-	-	9.082	9.856	8.665	9.727	9.726	9.411

Table 8: **Marathi** - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
64	2	4.46	1.514	6.663	7.097	7.469	7.797	8.424	7.490
64	6	4.65	1.245	6.543	7.225	7.482	7.975	8.454	7.544
64	12	5.00	1.136	6.760	7.289	7.563	7.968	8.507	7.617
512	2	41.00	0.693	7.373	7.494	7.644	8.314	8.782	7.922
512	6	54.00	0.569	7.507	7.645	7.693	8.420	8.816	8.016
512	12	73.00	0.544	7.525	7.718	7.743	8.450	8.836	8.054
1024	2	94.00	0.609	7.407	7.470	7.626	8.293	8.786	7.916
1024	7	157.00	0.557	7.567	7.639	7.740	8.409	8.832	8.037
GPT-4o-mini	-	-	-	8.953	9.692	8.637	9.695	9.734	9.342
GPT-4	-	-	-	9.340	9.786	8.800	9.821	9.827	9.515
Gemini-1.5-flash	-	-	-	9.283	9.920	8.944	9.910	9.904	9.592

Table 9: **Bengali** - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

A.2 Complete Tokenizer Evaluation Comparison

Tokenizer Name	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi							
Sarvam-1	0.518	7.734	7.783	7.806	8.554	8.912	8.158
SUTRA-mlt-256-v2	0.522	7.548	7.449	7.584	8.292	8.875	7.950
Tikoken	0.149	6.974	7.106	7.360	7.889	8.681	7.602
Marathi							
Sarvam	1.662	7.154	8.127	7.902	8.854	9.146	8.296
SUTRA	1.824	7.223	7.862	7.633	8.602	9.024	8.069
Tiktoken	1.167	6.514	7.442	7.437	8.105	8.851	7.670
Bangla							
Sarvam	0.569	7.507	7.645	7.693	8.420	8.816	8.016
SUTRA	0.608	7.614	7.374	7.595	8.212	8.845	7.928
Tiktoken	0.135	7.118	6.989	7.358	7.778	8.614	7.572

Table 10: **Tokenizer performance** across Hindi-Marathi-Bengali
Sarvam Tokenizer — 54M (E=512, L=6) Model — Lighter is better

B Inference Evaluation and Factual Comparison

B.1 SLM Inference and Evaluation Methodology

B.1.1 SLM Inference

Our SLMs are causal models, generating one token at a time, following a given input phrase. Generated stories are evaluated by GPT-4o using the prompt below. The overall score reported is the average of context awareness, completeness, grammar, fluency and creativity of the story.

Evaluation Prompt Template

```
{story}

The given {language} short story is
     for 5-7-year-old children.
    Keeping in mind the target
    demographic, rate the story on
    a scale of 1-10 for context
    awareness, completeness,
    grammar, fluency, and
    creativity. Evaluate context
    awareness by strictly assessing
     how well the story's middle
    and end align with the prompt
    "{prompt}". Also, provide an
    overall rating on a scale of
    1-10.

Only return a JSON dictionary in
    the following format:
{
"context awareness": "...",
"completeness": "...",
"grammar": "...",
"fluency": "...",
"creativity": "...",
"overall": "..."
}'
```

B.1.2 Evaluating an SLM

Using OpenAI's o1 model and subsequent manual refinement, we compiled a multilingual corpus of 1,000 matched prompts in Hindi, Marathi, and Bangla (see `training-inference/sample/`[2]). These prompts span 10 complementary category pairs to ensure broad evaluation coverage:

- Adventure & Fantasy
- Imagination & Creativity
- Curiosity & Discovery
- Mystery & Surprise
- Playfulness & Learning
- Family & Friendship
- Kindness & Happiness
- Helping & Sharing
- Courage & Perseverance
- Nature & Animals

For each prompt, the target model generates three distinct stories. In total, we evaluate (using GPT-4o) 3,000 narratives and compute the final performance metrics as the mean of their scores (see Appendix A).

B.2 Cross-Lingual Inference Comparison

In our evaluation, we compare inference performance across Hindi, Marathi, and Bangla using the 54 M model[3] with equivalent/matched prompts. Acknowledging the limitations discussed in Sec. G.1 and despite focusing on handpicked examples, we observe a consistent pattern of context drift in the Marathi outputs. As shown in Table 11, the Marathi story veers off to describe a "cave" and "stones" without ever mentioning the prompt's key term "riverbank," whereas the Hindi and Bangla versions remain faithful to the original context. This is further corroborated by the broader samples in Tables 13, 14, and 15. This behaviour is further analyzed and supported in Appendix G.10, where Marathi models of all sizes exhibit similar difficulties in preserving prompt relevance compared to their Hindi and Bangla counterparts.

B.3 Factual Prompt Evaluation

Evaluating our SLMs on short, factual prompts is essential for three reasons: (i) Probes their *context-tracking* capability. (ii) Tests *logical comprehension* rather than rote memorisation of training data and (iii) Reveals the models' propensity to *hallucinate* plausible but unsupported statements.

Using the multilingual factual benchmark in Table. 12, we assess multiple model sizes. The findings show a strong positive correlation between accurate context tracking and factual correctness—larger SLMs excel on both axes (cf. scores in A), whereas smaller models lag yet still display varied factual competence, indicating behaviour beyond simple memorisation (see Sec. G).

Among languages, Hindi models attain the highest accuracy. Surprisingly, Bangla models fall short of Marathi performance despite exhibiting superior context scores and manual inference traces, suggesting that Bangla might pose unique challenges to SLMs' logical reasoning.

[2] Available in our repository

[3] Sarvam tokenizer is used (see Section C.4)

.

Model	Completion	Evaluation
Hindi		
Prompt	**Indian Regional:** सूर्योदय पर, दो सबसे अच्छे दोस्त नदी के किनारे अजीब पैरों के निशान का पीछा कर रहे थे। **Translation:** At sunrise, two best friends were following some strange footprints along the riverbank.	
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** सूर्योदय पर, दो सबसे अच्छे दोस्त नदी के किनारे अजीब पैरों के निशान का पीछा कर रहे थे। जैसे ही सूरज की किरणें नदी पर पड़ीं, पानी में एक अद्भुत चमक दिखाई दी। सभी दोस्त एक-दूसरे को देखकर मुस्कुरा रहे थे। फिर अचानक, मोहन ने देखा कि पानी में कुछ रंग-बिरंगे फूल तैर रहे हैं। उसने सोचा, "अगर मैं इन फूलों को पकड़ लूं, तो यह बहुत मजेदार होगा!" उसने अपने दोस्तों को बुलाया और सबने मिलकर फूलों को पकड़ने की कोशिश की। खेल-खेल में, मोहन और उसके दोस्तों ने देखा कि नदी के किनारे एक छोटा सा रहस्य छिपा हुआ है। जब उन्होंने उन फूलों को पकड़ा, तो उन्हें पता चला कि वे सिर्फ सुंदर थे और उनमें जादुई शक्तियाँ थीं। सभी दोस्त उस दिन को कभी नहीं भूल पाए और हर सुबह सूर्योदय के समय फूलों के साथ खेलते रहे। **English Translation:** *At sunrise, two best friends were following some strange footprints along the riverbank.* As soon as the sun's rays touched the river, a magical sparkle appeared on the water. All the friends looked at each other and smiled. Suddenly, Mohan noticed some colorful flowers floating in the water. He thought, "If I could catch these flowers, it would be so much fun!" He called his friends, and together they tried to catch the flowers. While playing, Mohan and his friends discovered that a small secret was hidden along the riverbank. When they caught the flowers, they realized that they were not just beautiful—they had magical powers. None of the friends could ever forget that day, and from then on, they played with the flowers every morning at sunrise.	***Overall: 8.6*** *Context: 8* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*
Marathi		
Prompt	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. **Translation:** At sunrise, two best friends were following strange footprints by the river.	
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. त्यांनी एक गडद जागा पाहिली जिथे एक मोठा, चमकणारा दगड होता. "हे काय आहे?" समीर विचारला. आर्यनने उत्तर दिलं, "कदाचित हे नक्षत्र आहे. चला, आपण त्या दगडाच्या मागे जाऊन पाहू!" ते धाडसाने दगड ओलांडू लागले. दगडाच्या मागे एक लपलेलं गुहा दिसलं. गुहेच्या आत गेल्यावर, त्यांना अनेक रंगीत रत्न आणि चमकदार वस्त्रं सापडली. समीर आणि आर्यन खूप खूश झाले. त्यांनी ठरवलं की ते या गुप्त ठिकाणाला परत येतील. त्या दिवशी, त्यांनी खूप मजा केली आणि गुप्त ठिकाण शोधून काढले. त्यांनी एकत्र येऊन त्या गुप्त ठिकाणाबद्दल आपल्या मित्रांना सांगितलं. त्या दिवशी त्यांनी एक अद्भुत अनुभव घेतला, जो सदैव लक्षात राहील. **English Translation:** *At sunrise, two close friends were following unfamiliar footprints along the riverside.* They saw a dark place where there was a large, glowing stone. "What is this?" Sameer asked. Aryan replied, "Maybe it's a star fragment. Come on, let's go see behind that stone!" They bravely started crossing the stone. Behind the stone, they saw a hidden cave. When they entered the cave, they found many colorful gems and shining robes. Sameer and Aryan were very happy. They decided they would return to this secret place. That day, they had a lot of fun and discovered a hidden spot. Together, they told their friends about the secret place. That day, they had a wonderful experience that they would always remember.	***Overall: 8.2*** *Context: 6* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*
Bangla		
Prompt	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। **Translation:** At sunrise, two best friends were following strange footprints by the river.	
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। তাদের কাছে এক অদ্ভুত দারুন ছিল, তারা যেন তাদের জন্য নতুন একটি গল্প বলবে। এই গল্পটা ছিল খুব মজার এবং তারা সেই গল্প বলার সময় হাসাহাসি করতে লাগল। সন্ধ্যা হলে, তারা বাড়ির পথে হাঁটতে থাকল। সূর্য তখন আকাশে লাল হয়ে উঠেছিল এবং নদীর পাড়ে আধো আলোয় সবকিছুকে উজ্জ্বল করে দিয়েছিল। শিশুরা একসাথে হাঁটতে হাঁটতে মনে মনে ভাবছিল, আজকের দিনটা কত সুন্দর ছিল! তারা জানত, এই ধরনের সময়গুলো তাদের বন্ধুত্বকে আরও শক্তিশালী হবে। এখন তারা প্রতিদিন খেলতে আসবে, নতুন নতুন গল্প বলবে। নদীর শান্ত জল এবং সূর্যের আলোতে তাদের খেলা সব সময় মনে রাখার মতো থাকবে। **English Translation:** *At sunrise, two best friends were following strange footprints by the riverbank.* They had a strange charm with them, as if it would tell them a new story. That story was very funny and while telling it, they started laughing. In the evening, they began walking home. The sun had turned red in the sky and lit everything brightly along the riverbank in the dim light. The children, walking together, were thinking to themselves how beautiful the day was! They knew that moments like these would make their friendship even stronger. Now they will come to play every day, tell new stories. Their games in the calm water of the river and in the sunlight will always be something to remember.	***Overall: 8.6*** *Context: 9* *Fluency: 9* *Complete: 9* *Grammar: 8* *Creativity: 8*

Table 11: Hindi, Marathi & Bangla Inference comparisons across key model sizes.

Prompt ***Eng. Translation***	**5M** ***Hindi***	**54M** ***Hindi***	**157M** ***Hindi***	**5M** ***Marathi***	**54M** ***Marathi***	**5M** ***Bangla***	**54M** ***Bangla***
When Ram came back home, he was very tired, so he	was very happy. बहुत खुश था।	wanted to relax. आराम करना चाहता था।	he went to his bed to sleep. सो जाने के लिए अपनी बिस्तिर पर चला गया।	slept. झोपला.	went to sleep. झोपायला गेला.	sat down. বসে পড়ল।	went to sleep. সে ঘুমাতে গেল।।
Jack and Lily saw a rainbow after a rainy day. They were amazed by the colors. Jack said, "Look, Lily. A rainbow has	a lot of beauty! कितनी सुंदरता है!	seven colours! सात रंग हैं!	seven colours! सात रंग हैं!	a lot of beauty! किती सुंदरता आहे!	many colours किती रंग आहेत!	a lot of beauty! কত সুন্দর!	beautiful colours! একটি রং আছে!
Rahul wanted to read a book, so he	helped her friends to read. अपने दोस्तों को पढ़ने में मदद करती थी।	always studied first. हमेशा पहले पढ़ाई करता था।	could read new stories in it. नई-नई कहानियों को पढ़ सकता था।	was alone. एकटा होता	walked to his house. म्हणाला, "हे पुस्तक वाचण्यासाठी मला मदत हवी आहे"	got a new one. বইটি পড়তে শুরু করল।	decided that he would learn some new things. ঠিক করল যে সে কিছু নতুন জিনিস শিখবে।
If I throw a ball up in the air, eventually it will	fall! गिर गई!	fall! गिर गई!	fell into the pond. तालाब में गिर गई।	get stuck on a branch of a large tree. एका मोठ्या झाडाच्या फांदीवर अडकला	fly! उडेल!!	happen! ঘটবে!	happen! ঘটবে!
"Can cows fly?". Pooja asked her mother.	Mother said, "Son, this cow is so well made!" माँ ने कहा, "बेटा, यह गाय कितनी अच्छी बनी गई है!"	Her mother laughed and said, "No, cows are not meant to fly." उसकी माँ ने हंसते हुए कहा, "नहीं, गायें उड़ने के लिए नहीं हैं।"	Mother laughed and said, "If you enjoy cows flying, you can listen to them talk." माँ हंसते हुए बोली, "अगर तुम गायों को उड़ने का मजा लेते हो, तो तुम उनकी बातें सुन सकती हो।"	Her mother smiled and replied, "Yes, we'll take a cloth and let it fly!" आईने हसून उत्तर दिले, "हो, आपण एक कपडा घेऊ आणि त्याला उडू देऊ!"	Her mother smiled and said, "Sometimes, magic is all around us. You just have to believe." आईने हसून सांगितिले, "कधी कधी, जादू आपल्या आजुबाजूला असते. तुम्ही फक्त विश्वास ठेवा."	She said, "Mom, I will grow up to be a big dreamer." সে বলল, "মা, আমি বড় হয়ে একজন বড় স্বপ্ন দেখাবো।"	Mother smiled and said, "Yes, I will help you. If you can fly with me." মা হাসলেন এবং বললেন, "হ্যাঁ, আমি তোমাকে সাহায্য করব। তুমি যদি আমার সাথে উড়তে পারো।
It was winter and cold outside, so his mother told him, "You should	just rest a bit." बस थोड़ा आराम करना चाहिए।"	return home." तुम्हें घर लौटना चाहिए।"	go out wearing warm clothes." तुम्हें गर्म कपड़े पहनकर बाहर जाना चाहिए।"	~~should~~ are very nice. खूप चांगली आहेस.	give my friends hot hot tea!" माझ्या मित्रांना गरम गरम चहा देऊ शकते!"	~~You should~~ "If you work well, you can do anything." "তুমি যদি ভালোভাবে কাজ করো, তাহলে তুমি সবকিছু করতে পারবে।"	should be a little careful." "তুমি একটু সাবধান থাকতে হবে।"
"What do birds like to eat?" Tom asked his mother.	She said, "I will not get you a single fruit." उसने कहा, "मैं आपको एक फल का फल नहीं मिलाऊँगी।"	Mother said, "Pinky, we want fruits. माँ ने कहा, "पिकी, हमें फल चाहिए।"	Mother told him that she would bring him some fruits. माँ ने उसे बताया कि वह उसे कुछ फल लाकर देंगी।	He was thinking with his friends... तो मित्रांबरोबर विचार करत होता...	Mother told him, "The birds need some fruits to eat." आईने त्याला सांगितिले, "पक्ष्यांना खाण्यासाठी काही फळे लागतात. त्यामुळे त्यांना खायला मिळते."	They said Ronnie... তারা বলল, "রনি...	Mother said, "Birds eat our food, they sing." মা বলল, "পাখিরা আমাদের খাবার খায়, তারা গান গায়।"
If you put ice outside in the sun then	it will be very fun. यह बहुत मजेदार होगा.	it will melt and become water. यह पिघलकर पानी बन जाएगा।	it melts. पिघल जाती है.	it is very beautiful. ते खूप सुंदर आहे.	you will see something special! तुम्हाला एक विशेष गोष्ट दिसेल!	I will give you love. আমি তোমাকে ভালোবাসা দেবে।	you will see how beautiful it is! তাহলে তুমি দেখতে পাবে কত সুন্দর!
Human Evaluation	2.875 / 6	4.875 / 6	5.250 / 6	2.750 / 6	4.625 / 6	2.250 / 6	3.875 / 6

Table 12: Cross-lingual performance across key model sizes on **factual prompts** — *GPT-4 and better = 6/6*

Evaluation Scheme: Red = 1/6, Light Red = 2/6, Light Yellow = 3/6, Yellow = 4/6, Light Green = 5/6 & Green = 6/6

Left to Right, Red to Green, Arbitrary to Factual Contextual Completion

Model	Completion	Evaluation
Hindi		
Prompt	**Indian Regional:** घने जंगल में, एक चमकदार पेड़ रहस्यमय शक्ति से गूंज रहा था। **Translation:** In the dense forest, a glowing tree was echoing with mysterious power.	
Small *5M* *E = 64* *L = 12* *H = 8*	**Indian Regional:** घने जंगल में, एक चमकदार पेड़ रहस्यमय शक्ति से गूंज रहा था। एक दिन, उन्होंने देखा कि पेड़ पर बहुत सारे रंग-बिरंगे फूल खिले हुए हैं। मोहन ने सोचा, "ये फूल कितने सुंदर हैं!" उसने अपने दोस्तों से कहा, "चलो, हम यहाँ कुछ खास करते हैं।" सभी दोस्त एक साथ मिलकर पेड़ के नीचे बैठ गए। बच्चे बहुत खुश थे और उन्होंने तय किया कि वे यहाँ कुछ मजेदार करेंगे। मोहन ने एक छोटी सी बंदूक मांगी और कहा, "हम इसे पानी देंगे ताकि यह और भी सुंदर बने।" सभी ने मिलकर पत्थर को पानी दिया और बहुत ध्यान से काम किया। लेकिन एक दोस्त ने कहा, "तुम्हें इसे बहुत पसंद आएगा।" मोहन ने समझाया कि हमें एक-दूसरे के साथ मिलकर काम करना चाहिए। फिर, मोहन ने कहा, "हम सबको मिलकर काम करना चाहिए है। हमें हमेशा मिलकर काम करना चाहिए।" सभी ने सहमति जिताई और एक पेड़ के नीचे बैठ गए। जैसे ही उन्होंने घंटी को बजाया, पेड़ ने एक मीठी आवाज़ में कहा, "धन्यवाद, बच्चों! तुमने मेरे लिए यह कर दिया।" सभी बच्चे चौंक गए और पेड़ के पास बैठ गए। उस दिन मोहन और उसके दोस्तों ने सीखा कि एकता में ताकत होती है। उन्होंने समझा कि जब हम मिलकर काम करते हैं, तो हम हमेशा एक-दूसरे की मदद कर सकते हैं। और इस तरह, मोहन और उसके दोस्तों ने एक नई दोस्ती की शुरुआत की। **English Translation:** *In the dense forest, a glowing tree was echoing with mysterious power*. One day, they saw that many colorful flowers had bloomed on the tree. Mohan thought, "These flowers are so beautiful!" He said to his friends, "Come on, let's do something special here." All the friends sat together under the tree. The children were very happy and they decided that they would do something fun here. Mohan asked for a small gun and said, "We will give it water so that it becomes even more beautiful." Everyone gave water to the stone together and worked very carefully. But one friend said, "You will really like this." Mohan explained that we should all work together. Then Mohan said, "We all should work together. We should always work together." Everyone agreed and sat under a tree. As soon as they rang the bell, the tree said in a sweet voice, "Thank you, children! You did this for me." All the children were surprised and sat near the tree. That day, Mohan and his friends learned that there is strength in unity. They understood that when we work together, we can always help each other. And in this way, Mohan and his friends began a new friendship.	***Overall: 7.6*** *Context: 6* *Fluency: 8* *Complete: 8* *Grammar: 9* *Creativity: 7*
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** घने जंगल में, एक चमकदार पेड़ रहस्यमय शक्ति से गूंज रहा था। पेड़ की पत्तियाँ सूरज की रोशनी में चमक रही थीं। बच्चे उस पेड़ के पास गए और देखा कि पेड़ के नीचे एक प्यारा सा खरगोश बैठा है। खरगोश ने कहा, "मैंने सुना है कि इस जंगल में एक जादुई फूल है, जो हर साल एक बार खिलता है, तुम्हारी इच्छा पूरी करता है।" बच्चों ने सोचा कि अगर वे उस जादुई फूल को ढूंढ लेंगे, तो उनकी सभी इच्छाएँ पूरी हो जाएँगी। उन्होंने खरगोश से पूछा, "हम उस फूल को कैसे ढूंढ सकते हैं?" खरगोश ने मुस्कुराते हुए कहा, "तुम्हें अपने दिल की सच्चाई से काम करना होगा।" बच्चों ने मिलकर जंगल में खोज शुरू की। उन्होंने कई पेड़ और झाड़ियाँ पार कीं। अंत में, उन्होंने एक चमकदार पेड़ देखा, जो पहले से भी ज्यादा सुंदर था। बच्चों ने उस पेड़ के चारों ओर बैठकर अपनी इच्छाएँ मांगी। अचानक, पेड़ पर एक सुगंधित फूल खिलता हुआ दिखाई दिया। बच्चों ने उस फूल को देखकर एक-दूसरे को देखा और उनके दिलों में एक रहस्यमय रूप से गरमाहट भरा अनुभव हुआ। **English Translation:** *In the dense forest, a glowing tree was echoing with mysterious power*. The leaves of the tree were shining in the sunlight. The children went near that tree and saw that a cute little rabbit was sitting under the tree. The rabbit said, "I have heard that there is a magical flower in this forest, which blooms only once a year and fulfills your wish." The children thought that if they found that magical flower, all their wishes would come true. They asked the rabbit, "How can we find that flower?" The rabbit smiled and said, "You must work with the truth of your heart." The children together began searching in the forest. They crossed many trees and bushes. Finally, they saw a glowing tree, even more beautiful than before. The children sat around that tree and made their wishes. Suddenly, a fragrant flower bloomed on the tree. Seeing that flower, the children looked at each other and felt a mysteriously warm sensation in their hearts.	***Overall: 8.8*** *Context: 9* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*
Large *157M* *E = 1024* *L = 7* *H = 8*	**Indian Regional:** घने जंगल में, एक चमकदार पेड़ रहस्यमय शक्ति से गूंज रहा था। सभी जानवर उस पेड़ के चारों ओर इकट्ठा हो रहे थे। अचानक, पेड़ से एक सुनहरी रोशनी निकली और जंगल में गर्माहट फैल गई। सभी जानवरों ने देखा कि पेड़ के पास एक सुंदर परी प्रकट हुई। परी ने कहा, "तुम सबका स्वागत है! आज का दिन बहुत खास है।" परी ने सभी जानवरों को एक खेल खेलने के लिए कहा। उन्होंने मिलकर एक खेल खेला जिसमें उन्हें छिपना और ढूँढना था। जैसे ही खेल शुरू हुआ, जंगल में एक अद्भुत संगीत गूंजने लगा। सभी जानवर छिपने के लिए दौड़ पड़े। चिड़ियों ने मीठी आवाज़ में गाना शुरू किया और खरगोश ने अपने लंबे कान छुपने के स्थान से सबको देख लिया। जब खेल खत्म हुआ, तो सभी जानवर बहुत खुश थे। परी ने कहा, "तुम सबने बहुत अच्छा खेला। याद रखो, दोस्ती और खेल से जीवन में हमेशा गरमाहट भरी रहती है।" फिर परी, उस रहस्यमय पेड़ के पास सभी जानवरों ने एक-दूसरे को गले लगाया और खुशी से झूम उठे। फिर परी ने अपनी जादुई शक्तियों से सभी जानवरों को एक खास उपहार दिया। **English Translation:** *In the dense forest, a glowing tree resonated with mysterious power*. All the animals were gathering around the tree. Suddenly, a golden light emerged from the tree, spreading warmth throughout the forest. All the animals saw a beautiful fairy appear near the tree. The fairy said, "Welcome, everyone! Today is a very special day." The fairy asked all the animals to play a game. Together, they played a game of hide and seek. As soon as the game began, a wonderful melody echoed through the forest. All the animals ran to hide. The birds began to sing in sweet voices, and the rabbit, with its long ears, watched everyone from its hiding spot. When the game ended, all the animals were very happy. The fairy said, "You all played very well. Remember, life is always full of warmth through friendship and play." Then, near the mysterious tree, all the animals hugged each other and danced with joy. After that, the fairy used her magical powers to give each animal a special gift.	***Overall: 9.6*** *Context: 9* *Fluency: 10* *Complete: 10* *Grammar: 10* *Creativity: 9*

Table 13: **Hindi Inference** comparisons across key model sizes.

Model	Completion	Evaluation
Marathi		
Prompt	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. **Translation:** At sunrise, two close friends were following unfamiliar footprints by the riverside.	
Small *5M* *E = 64* *L = 12* *H = 8*	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. त्यांनी एकत्र येऊन एक मोठा झाड पाहिला, जो त्यांना पाहून त्यांच्यातील आवाज ऐश्वर्यपूर्ण आवाजात एक अद्भुत दृश्य दिसलं. ते झाड खूप सुंदर होते. आर्यन त्या झाडाच्या पानांवर चढून खेळत होता. त्यांच्या आवाज आणि आर्यनने ठरवलं की, ते झाडावर चढून, झाडावर चढतील. झाडावर चढल्यानंतर आर्यनने एका मोठ्या झाडाच्या खाली एक मोठा पक्षी पाहिला. तो पक्षी आकाशात उडत होता, पण आर्यन खूप खूष झाला. आर्यनने त्या पक्षाला स्पर्श केला आणि त्याच्या पानावर एक मोठा आवाज आला. त्या दिवशी आर्यनने त्या पक्ष्याने एक जादुई पंख घेतला आणि त्या पक्ष्याला एकत्र बसले. ते पक्षी आकाशात उडायला लागले, पण चिरतन एकदम सुंदर होता. तो पक्षी खूप सुंदर आणि सुंदर होता, तो त्याला म्हणाला, "तू एकटा असणारा गाणं गाणं ऐकतो, परंतु मी तुझ्या सोबत आहेस!" आर्यनने हळू हळू हळूच उचलून त्याच्या गाण्यात रमले एक अद्भुत गोष्ट शिकली. आता तो एकदम शांत होता, जो सर्वांच्या मनात राहिला. **English Translation:** *At sunrise, two close friends were following unfamiliar footprints by the riverside*. Together, they came upon a large tree, and as they looked at it, a majestic voice echoed within them, revealing a wondrous sight. The tree was incredibly beautiful. Aryan began climbing and playing among its leaves. Moved by the voice and their curiosity, Aryan decided to climb the tree. Once at the top, he saw a large bird resting beneath the tree. The bird soon soared into the sky, and Aryan felt immense joy. As he reached out and touched the bird, a loud sound came from the leaves. That day, Aryan received a magical feather from the bird, and he sat with the bird. The bird began flying through the sky, and the moment was breathtaking. The bird was stunning and graceful. It said to Aryan, "You may sing songs alone, but I am with you!" Slowly, Aryan lifted his spirits and became immersed in its song, learning a wonderful lesson. From that moment on, he was filled with peace—a peace that stayed in everyone's hearts.	***Overall: 6.6*** *Context: 6* *Fluency: 6* *Complete: 7* *Grammar: 6* *Creativity: 8*
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. त्यांनी एक गडद जागा पाहिली जिथ एक मोठा, चमकणारा दगड होता. "हे काय आहे?" समीर विचारला. आर्यनने उत्तर दिलं, "कदाचित हे नक्षत्र आहे. चला, आपण त्या दगडाच्या मागे जाऊन पाहू!" ते धाडसाने दगड ओलांडू लागले. दगडाच्या मागे एक लपलेलं गुहा दिसलं. गुहेच्या आत गेल्यावर, त्यांना अनेक रंगीत रत्न आणि चमकदार वस्त्रं सापडली. समीर आणि आर्यन खूप खूश झाले. त्यांनी ठरवलं की ते या गुप्त ठिकाणाला परत येतील. त्या दिवशी, त्यांनी खूप मजा केली आणि गुप्त ठिकाण शोधून काढले. त्यांनी एकत्र येऊन त्या गुप्त ठिकाणाबद्दल आपल्या मित्रांना सांगितलं. त्या दिवशी त्यांनी एक अद्भुत अनुभव घेतला, जो सदैव लक्षात राहील. **English Translation:** *At sunrise, two close friends were following unfamiliar footprints along the riverside.* They saw a dark place where there was a large, glowing stone. "What is this?" Sameer asked. Aryan replied, "Maybe it's a star fragment. Come on, let's go see behind that stone!" They bravely started crossing the stone. Behind the stone, they saw a hidden cave. When they entered the cave, they found many colorful gems and shining robes. Sameer and Aryan were very happy. They decided they would return to this secret place. That day, they had a lot of fun and discovered a hidden spot. Together, they told their friends about the secret place. That day, they had a wonderful experience that they would always remember.	***Overall: 8.2*** *Context: 6* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*
Large *157M* *E = 1024* *L = 7* *H = 8*	**Indian Regional:** सूर्योदयाच्या वेळी, दोन घट्ट मित्र नदीकाठच्या अनोळखी पाऊलखुणांचा पाठलाग करत होते. ते एका ठिकाणी गेल्यावर त्यांनी एक सुंदर फुलांचा बाग शोधला. बागेत रंग-बिरंगी फुलं होती, जी सर्वत्र पसरली होती. नवीन मित्रांनी ठरवलं की ते या बागेत खेळतील. त्यांनी एकत्रीत येऊन एक छान खेळ खेळण्यासाठी ठरवलं. पण त्यांच्या मनात एक गोष्ट होती, की त्यांनी या बागेतील फुलांचा सुगंध साठवणे आवश्यक आहे. त्यामुळे त्यांनी फुलांचे तुकडे तोडले आणि आपल्या घरात नेले. जसजसे वेळ जात होता, त्यांच्या घरात सुगंध पसरला. सर्व गप्पा, हसणं आणि खेळणं यामुळे त्यांच्या मित्रत्वाच्या बंधनात आणखी एक सुंदर रंग भरला. त्यांना समजलं की, फुलांचे सौंदर्य त्यांच्या मित्रत्वात आहे. या अनुभवाने त्यांच्या मनात एक खास आठवण राहिली, जी सदैव त्यांच्या हृदयात राहील **English Translation:** *At sunrise, two close friends were following some unfamiliar footprints along the riverbank*. As they reached a certain place, they discovered a beautiful flower garden. The garden was filled with colorful flowers spread all around. The new friends decided that they would play in this garden. They came together and agreed to play a nice game. But one thought stayed in their minds — they needed to preserve the fragrance of the flowers in the garden. So, they picked a few flower petals and took them home. As time passed, their home was filled with a sweet fragrance. All the conversations, laughter, and play added another beautiful color to the bond of their friendship. They realized that the beauty of the flowers lived on in their friendship. This experience left a special memory in their hearts, one that would remain with them forever.	***Overall: 8.2*** *Context: 7* *Fluency: 8* *Complete: 9* *Grammar: 9* *Creativity: 8*

Table 14: **Marathi Inference** comparisons across key model sizes.

Model	Completion	Evaluation
Bangla		
Prompt	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। **Translation:** At sunrise, two best friends were following strange footprints by the river.	
Small *5M* *E = 64* *L = 12* *H = 8*	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। নদীটি খুব সুন্দর, রঙিন এবং অনেক মাছ খরিও আছে। রাহুল বলল, "দেখো, সেই মাছগুলো কেমন অদ্ভুত!" সবাই মিলে মাছগুলো দেখছিল এবং আনন্দে চিৎকার করছিল। হঠাৎ, রাহুলের এক বন্ধু বলল, "চলো, আমরা মাছ ধরতে যাই!" তারা সবাই নদীর কাছে গেল। রাহুল খুব উত্তেজিত হয়ে উঠল। সে ভাবল, "যদি আমি মাছ ধরতে পারি, তাহলে আমি আমার মায়ের জন্য একটি বড় মাছ বানাবো!" কিন্তু মাছ ধরার সময়, রাহুলের জাল খুব একটা সফল হল না। তবে, রাহুল হাল ছাড়ল না। সে আবার চেষ্টা করল। সে ধীরে ধীরে জালটি পানিতে ফেলল এবং এইবার একটি বড় মাছ ধরতে সক্ষম হল! তার বন্ধুদের মধ্যে আনন্দের ঢেউ উঠল। তারা সবাই রাহুলকে অভিনন্দন জানাল। রাহুল বুঝতে পারল, চেষ্টা করলে সবকিছুই সম্ভব। সে তার মায়ের জন্য একটি বড় মাছ নিয়ে বাড়ি ফিরিল। তার মনে নতুন আশা জাগল, যে সে ভবিষ্যতে আরও অনেক কিছু করতে পারবে। **English Translation:** *At sunrise, two best friends were following strange footprints by the river.* The river was very beautiful, colorful, and filled with many fish. Rahul said, "Look, how strange those fish are!" Everyone watched the fish together and shouted in joy. Suddenly, one of Rahul's friends said, "Let's go fishing!" They all went near the river. Rahul became very excited. He thought, "If I can catch a fish, I will cook a big one for my mother!" But while fishing, Rahul's net didn't work very well. Still, Rahul didn't give up. He tried again. Slowly, he cast his net into the water, and this time he managed to catch a big fish! A wave of joy spread among his friends. They all congratulated Rahul. Rahul realized that anything is possible if you try. He returned home with the big fish for his mother. A new hope awakened in his heart—that he could achieve many more things in the future.	***Overall: 8.2*** *Context: 6* *Fluency: 9* *Complete: 9* *Grammar: 9* *Creativity: 8*
Medium *54M* *E = 512* *L = 6* *H = 8*	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। তাদের কাছে এক অদ্ভুত দারুন ছিল, তারা যেন তাদের জন্য নতুন একটি গল্প বলবে। এই গল্পটা ছিল খুব মজার এবং তারা সেই গল্প বলার সময় হাসাহাসি করতে লাগল। সন্ধ্যা হলে, তারা বাড়ির পথে হাঁটতে থাকল। সূর্য তখন আকাশে লাল হয়ে উঠেছিল এবং নদীর পাড়ে আধো আলোয় সবকিছুকে উজ্জ্বল করে দিয়েছিল। শিশুরা একসাথে হাঁটতে হাঁটতে মনে মনে ভাবছিল, আজকের দিনটা কত সুন্দর ছিল! তারা জানত, এই ধরনের সময়গুলো তাদের বন্ধুত্বকে আরও শক্তিশালী হবে। এখন তারা প্রতিদিন খেলতে আসবে, নতুন নতুন গল্প বলবে। নদীর শান্ত জল এবং সূর্যের আলোতে তাদের খেলা সব সময় মনে রাখার মতো থাকবে। **English Translation:** *At sunrise, two best friends were following strange footprints by the riverbank.* They had a strange charm with them, as if it would tell them a new story. That story was very funny and while telling it, they started laughing. In the evening, they began walking home. The sun had turned red in the sky and lit everything brightly along the riverbank in the dim light. The children, walking together, were thinking to themselves how beautiful the day was! They knew that moments like these would make their friendship even stronger. Now they will come to play every day, tell new stories. Their games in the calm water of the river and in the sunlight will always be something to remember.	***Overall: 8.6*** *Context: 9* *Fluency: 9* *Complete: 9* *Grammar: 8* *Creativity: 8*
Large *157M* *E = 1024* *L = 7* *H = 8*	**Indian Regional:** সূর্যোদয়ের সময়, দুই সেরা বন্ধু নদীর ধারে অদ্ভুত পায়ের ছাপ অনুসরণ করছিল। সে মনে মনে ঠিক করলো, আজকে একটু দূরে যাবার দেখা হবে। তাই তারা একটি বড় পাথরের পিছনে গেল। হঠাৎ, তারা একটি ছোট মাছ দেখতে পেলো, যা পানির মধ্যে ঝলমল করছিল। মাছটি ছিল দিব্য রঙের। রাহুল আর সুমি খুব আনন্দিত হলো। তারা মাছটির দিকে নজর দিলো আরো কিছু সময় ধরে মাছটির নাচার জন্য চেষ্টা করতে লাগলো। নদীর শান্ত জল তাদের আনন্দে ভরে দিলো। শেষে, মাছটি তাদের কাছে এসে একবার ঝলমল করে উঠলো। রাহুল আর সুমি হাসতে হাসতে বললো, "ওহ, কি সুন্দর!" তারা বুঝলো, মাছটি তাদের বন্ধু হয়ে গেছে। এরপর তারা নদীর ধারে বসে গল্প করতে লাগলো। সেই দিনটি তাদের মনে রয়ে গেলো, যেন নদীর শান্ত জল আর মাছটির নাচার ছোঁয়া। **English Translation:** *At sunrise, two best friends were following strange footprints along the riverbank.* They decided to meet up a little further away today. So they went behind a large rock. Suddenly, they saw a small fish shimmering in the water. The fish was brilliantly colored. Rahul and Sumi were overjoyed. They kept their eyes on the fish and tried for some time to make it dance. The calm river water filled them with joy. Finally, the fish came close to them and sparkled once more. Rahul and Sumi laughed and said, "Oh, how beautiful!" They realized that the fish had become their friend. After that, they sat by the riverside and started chatting. That day stayed in their memory, as if touched by the calm river water and the fish's dance..	***Overall: 8.2*** *Context: 7* *Fluency: 8* *Complete: 9* *Grammar: 9* *Creativity: 8*

Table 15: **Bangla Inference** comparisons across key model sizes.

C SLM Architecture & Training, Tokenizer Details

The implementation for this section is available at `training-inference/` in our repository.

C.1 nanoGPT based SLM Architecture

We build all RegionalTinyStories models on **nanoGPT** (Karpathy, 2022), a 100,% PyTorch re-implementation of GPT-2 that exposes the full training loop in $\sim$ 300 lines of code.

The core module is a decoder-only Transformer with pre-norm residual blocks:

$$y = x + \text{MHA}(x), \tag{1}$$

$$z = y + \text{MLP}(\text{LN}(y)), \tag{2}$$

$$\text{Block}(x) = \text{LN}(z). \tag{3}$$

We first apply multi-head attention (MHA) with a residual (y), feed the result through a feed-forward network with another residual (z), and finally apply layer normalisation (LN); this is the standard pre-norm Transformer block used in nanoGPT. MHA uses *flash-attention-2* kernels for memory-efficient training, and the MLP employs a GELU activation.

Model grid We sweep over the following hyper-parameters to obtain **8**, **19**, **54**, **73**, and **153** M-parameter checkpoints per language:

- *Embedding dim* $\in \{64, 256, 512, 768, 1024\}$,
- *Layers* $\in \{2, 6, 12\}$ (7 for the 153M model),
- *Heads* $= 8$ (fixed across all sizes),
- *Context length* $= 1024$ tokens.

Tokenisation Each language is pre-tokenised with **Sarvam** (main experiments), **SUTRA**, or **Tiktoken**. (Refer C.4 for details on tokenizers and the number of tokens in our datasets.) The resulting token ID streams are packed into a contiguous list and written to `train.bin`/`val.bin`, expected by nanoGPT.

Configurations

- *Config. common to all models:*
  - **Optimizer**: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, λ=0.1).
  - **Total Epochs**: 5000
  - **Minimum LR**: Cosine decay, after Warmup Epochs (below) from Base LR (below) to 6×10^{-5}.
  - **Precision**: FP16 (CPU) / bfloat16 (CUDA); Gradient Clipping at 1.0.
  - **Regularisation**: Dropout 0.0 (OFF) on attention, MLP, and embeddings.
- *Config. unique to each model:*
  - For 157M (largest) model:
    - **Base LR**: 8×10^{-4}
    - **Wamrup Epochs**: 450
    - **Batching**: 96 sequences ($1024 \times 98 = 100,352$ tokens); 40 gradient accumulation steps.
    - **Embedding/Hidden dim.**: 1024
    - **Attention Layers**: 7
    - **Attention Heads**: 8

Reproducibility Training configurations can be found and easily customised through (`training-inference/config.py`) found in our repository.

C.2 Training Details

Training Regime

- **Training Epochs**: 5000 (~convergence)
- **Testing Epochs**: 50
- **Testing Frequency**: 200 epochs
- **Logging Frequency**: 2 epochs

Hardware Details We utilise a DDP for multi-GPU training, where each GPU randomly samples a batch from the training/testing data. We observe, the total training time is only affected by the total combined VRAM (of all GPUs).

Model Size	Training Time	Cost ($2.0/hr)
5M	~6 hr	~12 USD
54M	~8 hr	~16 USD
157M	~16hr	~32 USD

Table 16: **Training time (on 1 x H100) and Cost** for different model sizes. *Metrics are similar across languages.*

Optimal Hardware Using 2×H100 doubles both VRAM and hourly cost but halves training time, keeping total cost unchanged. **As training is VRAM-bound rather than FLOP- or architecture-dependent, the RTX A6000 offers the best cost efficiency per GB of VRAM.**

C.3 Training (WANDB) Logs

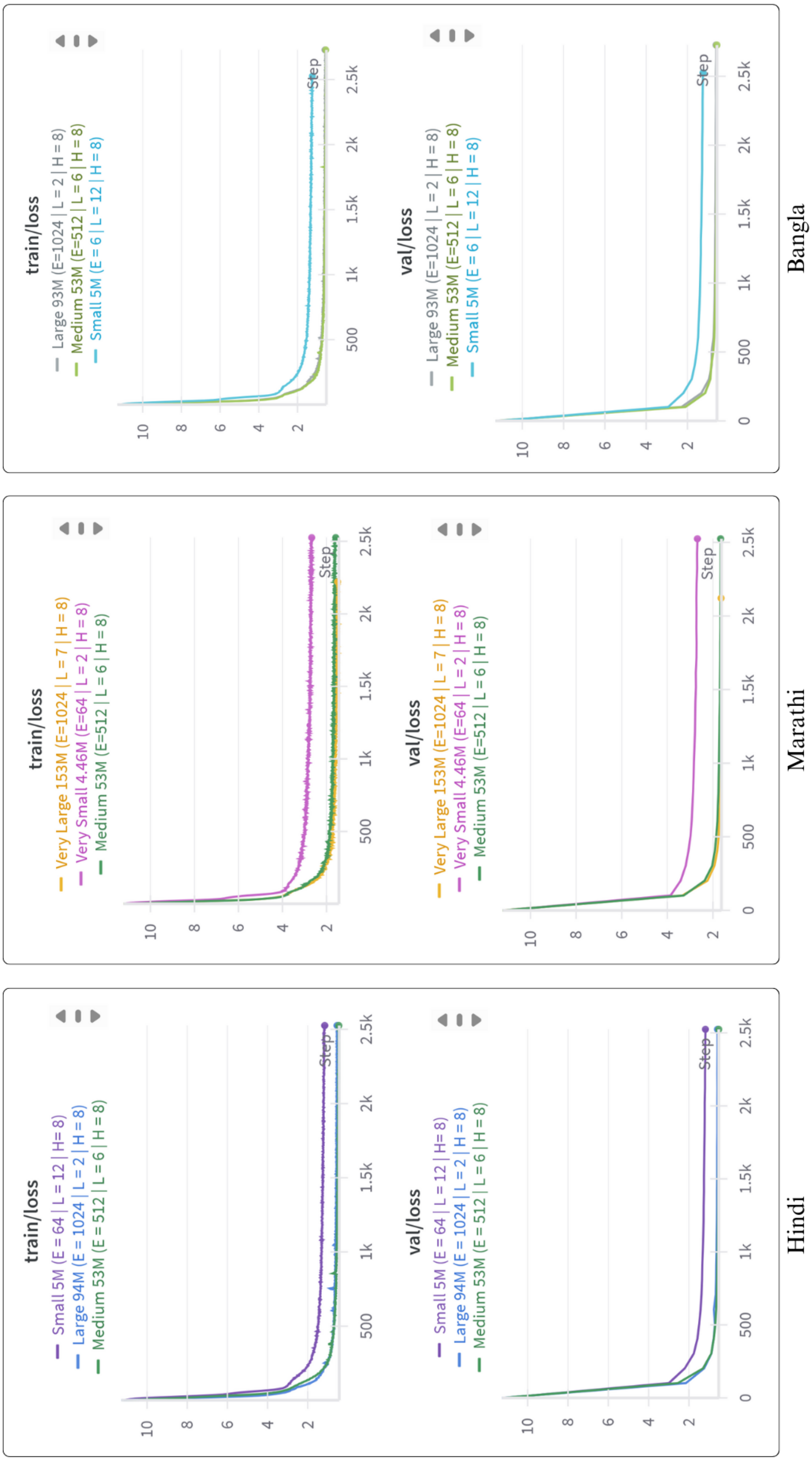

Figure 2: Training & Testing loss curves for Small, Medium and Large Models across Hindi, Marathi & Bangla

C.4 Tokenizer Details

Tokenizer	Hindi	Marathi	Bangla
Sarvam	~658 M	~639 M	~618 M
SUTRA	~598 M	~566 M	~542 M
Tiktoken	~1.3 B	~1.1 B	~0.96 B

Table 17: Tokenizer-wise Token Distribution Across Languages

Vocabulary Size

- Sarvam-1: 68096
- SUTRA-mlt-256-v2: 50304
- Tikoken: 256064

Underlying Framework

- Sarvam-1: SentencePiece
- SUTRA-mlt-256-v2: SentencePiece
- Tikoken: Byte-Pair Encoding

Sarvam–the ideal choice As we show in Table. 10 the Sarvam-1 tokenizer outperforms SUTRA-mlt-256-v2 across all metrics. Additionally, Sarvam has a faster tokenisation speed and, more importantly, has a significantly smaller vocabulary size. Subsequently, models trained using the Sarvam tokenizer have substantially smaller total parameters (owing to a smaller Token-embedding matrix and Output-projection matrix). For these reasons, we utilise Sarvam-1 as our primary tokenizer for all analysis unless specified otherwise. For a more detailed analysis of the 3 tokenizers refer Sec. H.

D Analysis of Translated Dataset & Comparative Evaluation of Synthetic Versus Translated Data

D.1 Translation Methodology

The original TinyStories dataset in English (Eldan and Li, 2023b) has a train split that contains 2,119,719 stories, with 320,470 duplicates, reducing unique stories to 1,799,249 (15.12% duplicates). This may impact model training, potentially skewing results.

Examples of stories appearing three times:

- ”Lily and Ben were playing in the park. They saw a fox hiding in the bushes. The fox had red fur and ...”
- ”Tom is a boy who likes to play with his lab. His lab is a big, black dog who can do amazing tricks. ...”
- ”Sara loved her pet cat, Lily. Lily was soft and fluffy and liked to sleep on Sara’s bed. Sara liked ...”
- ”Sam was sick. He had a bad cough and a sore throat. He did not like being sick. He wanted to play wi...”

Validation split has 21,990 unique stories without duplicates, ensuring reliable evaluation. However, cross-split analysis shows 6,601 stories appear in both splits, which could inflate model performance metrics. Addressing these duplicates is crucial for accurate model comparison between translated and synthetic data, enhancing the robustness of our findings in regional language modeling. Hence, we removed the duplicates and merged the two datasets.

It was decided to use GPT4o for rating the translation quality based on recent reports (Kocmi and Federmann, 2023; Jiao et al., 2023) of GPT-4 matching or outperforming Google Translate in the case of non-English languages. In the context of evaluating translation models for Hindi and Bangla, using GPT-4o, several multilingual models were assessed to identify the most effective translation method. The models include mBART (Liu et al., 2020), IndicTrans2 (Gala et al., 2023), mT5 (Xue et al., 2020), Helsinki-NLP OPUS MT (Tiedemann et al., 2023), IndicBART (Dabre et al., 2021), NLLB (Team et al., 2022), and M2M100 (Fan et al., 2021). Each model offers distinct capabilities, particularly in handling translations between English and Indic languages.

GPT-4o was given the original English text and asked to translate it into the target Indic language. Following this, it was provided with one of the machine translations and asked to rate it out of 10, using the following prompt :

I have used a machine translation model to translate the original story. On a scale of 1-10, evaluate the translation quality of the story, with respect to your translation, 1 being very bad and 10 being of the same quality as yours. Remember that ”quality” here does not mean the same words, but meaningfully retaining the same context and fluency. Also, point out each instance where there is a mistake in translation.

For Bangla translation, IndicTrans2 achieved an average score of 7/10 on 100 stories. Google Translate, although widely used, sometimes produces translations with contextual inaccuracies and verb errors. Still, it was rated at an average score of 8 by GPT-4o. Hence, it was chosen for Bangla. On the other hand, the NLLB model received an average score of 7/10. GPT-4o noted that while the NLLB translations were generally good, they lacked the fluency and natural phrasing found in Google Translate. Specific issues were identified with verb tenses and word choices, leading to minor awkwardness in sentence structures.

However, for Hindi translations, both NLLB and Google Translate produced scores around 8.5. Hence, the total Hindi translation was split between these two, primarily due to resource planning regarding API calls for Google Translate and GPU time for NLLB.

Overall, these evaluations demonstrate the strengths and weaknesses of different models in translating between English and Indic languages, highlighting the importance of choosing a model that balances semantic accuracy with natural language fluency. This analysis is critical for understanding the challenges and opportunities in model translation quality, particularly in academic contexts where precise language use is essential.

D.2 Synthetic versus Translated Data

Trained On	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi							
Synthetic Data	0.518	7.734	7.783	7.806	8.554	8.912	8.158
Translated Data	1.385	5.969	5.551	5.742	6.638	7.692	6.298
Marathi							
Synthetic Data	1.662	7.154	8.127	7.902	8.854	9.146	8.296
Translated Data	2.524	6.423	6.932	6.431	7.312	8.218	7.063
Bangla							
Synthetic Data	0.569	7.507	7.645	7.693	8.420	8.816	8.016
Translated Data	1.494	6.879	6.599	6.462	7.340	8.122	7.080

Table 18: Performance comparison of **Models trained on Translated vs Synthetic Data** performance across Hindi-Marathi-Bangla. *54M (E=512, L=6) Model — Lighter is better*

Training solely on our carefully generated synthetic corpus significantly boosts model performance across all metrics, across the evaluated languages (Table 18).

Hindi translated data trails that of Marathi and Bangla, likely because the Hindi set was produced with two independent translation pipelines that introduce additional noise.

Additionally, Boughorbel et al.(2024) propose an orthogonal solution: they inject a small amount of high-quality synthetic data into MT-generated corpora to cut cultural/linguistic bias in an Arabic story-generation task, as evidenced by changes in sparse auto-encoder features. Their results suggest that selective fine-tuning, rather than fully synthetic training, can also improve models trained on translated data.

E Synthetic Dataset Generation through LLM Prompting

E.1 Prompt Generation Methodology

Building upon the seminal work of TinyStories (Eldan and Li, 2023b), the prompt generation process began with creating comprehensive lexical resources for each target language: Hindi, Bangla, and Marathi. We compiled vocabulary lists consisting of approximately 300 nouns, 300 verbs, and 300 adjectives appropriate for children aged 5-7 years for each of the languages. These were stored in language-specific text files.

Additionally, we developed "features" lists in both English and the target languages. Original TinyStories contains some features inappropriate for children (violence, unhappy endings, etc.). We improve on this by ensuring our features represent positive, age-appropriate narrative elements, themes, or tones to guide story generation (e.g., learning values, friendship themes, acts of kindness). These resources were consolidated into a structured JSON format for each language.

E.2 Unique Prompt Generation Algorithm

The foundational TinyStories paper fails to mention the methodology used to generate prompts, we observe a high frequency of repetitions in their dataset as described in section D.1. To ensure maximum diversity in the dataset while preventing duplicates, we implemented Alg. 1.

This approach effectively prevents repetition patterns (two prompts with the same noun, verb, adjective and feature, & two prompts with the same noun, verb, adjective) in the dataset, eliminating approximately 250,000 (on average) potential duplicate prompts from the target 3M dataset per language. The tracking of both quadruplet and triplet identifiers ensured maximum lexical diversity in the stories.

E.3 Prompt Complexity Evolution

The seminal work does not justify its choice of prompt. To identify the optimal configuration for generating high-quality children's stories, we systematically evaluate different prompt complexity levels.

Five distinct complexity levels, with increasing sophistication, were developed and evaluated:

- **Level 1**: Basic structure (TinyStories baseline) with minimal guidance
- **Level 2**: Enhanced structure with explicit narrative guidance (beginning/middle/end) and tone constraints
- **Level 2+**: Extended word limit (350-500 words) while maintaining structural guidance
- **Level 3**: Addition of dialogue elements (maximum three exchanges) and thematic guidance
- **Level 4**: Incorporation of cultural references (e.g., Panchatantra, Tenali Raman stories)
- **Level 4+/5**: Extension with supporting characters and natural elements

Algorithm 1 Unique Prompt Generation

$UsedIDs \leftarrow \emptyset$
$UsedTriplets \leftarrow \emptyset$
$Prompts \leftarrow \emptyset$
$DuplicateCount \leftarrow 0$
while $|Prompts| < TargetCount$ **do**
  $n \leftarrow$ Select random element from N_{words}
  $v \leftarrow$ Select random element from V_{words}
  $a \leftarrow$ Select random element from A_{words}
  $f \leftarrow$ Select random element from F_{words}
  $ID \leftarrow \text{ConcatenateIndices}(n, v, a, f)$
  $TripletID \leftarrow \text{ConcatenateIndices}(n, v, a)$
  if $ID \notin UsedIDs$ and $TripletID \notin UsedTriplets$ **then**
    $UsedIDs \leftarrow UsedIDs \cup \{ID\}$
    $UsedTriplets \leftarrow UsedTriplets \cup \{TripletID\}$
    $prompt \leftarrow \text{FormatTemplate}(n, v, a, f)$
    $Prompts \leftarrow Prompts \cup \{prompt\}$
  else
    $DuplicateCount \leftarrow DuplicateCount + 1$
  end if
end while
return $Prompts, DuplicateCount$

Using each prompt template, we ask GPT-4o-mini to generate 1000 stories. These 1000 stories are then evaluated by GPT-4o; these results (Fig. 3) are used to determine the optimal prompt (below Fig. 4).

E.4 Optimal Prompt Template

Based on the aforementioned cross-prompt evaluation (Fig. 3), the Level 2+ template, which produced the best results across languages, followed the structure below:

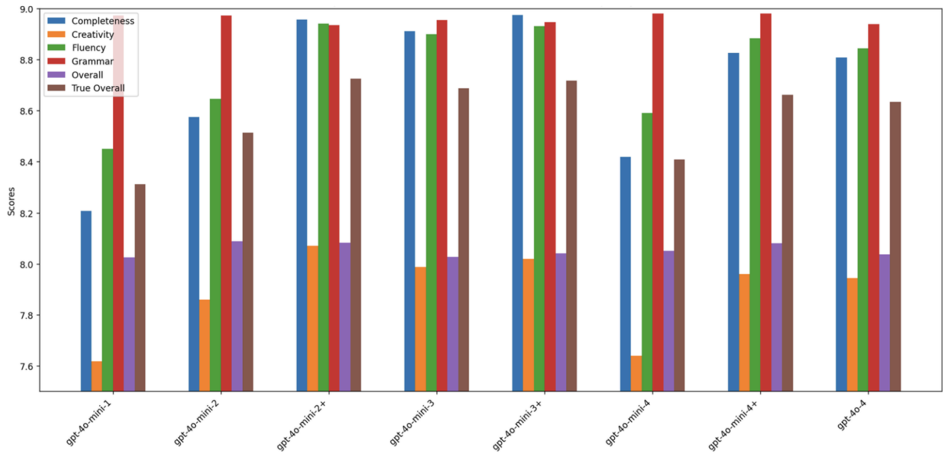

Figure 3: **Prompt Complexity Comparison**
gpt-4o-mini-1: Model used = GPT-4o-mini, Prompt complexity = 1
Prompt complexity = 2+ (where + indicates increased word limit of 350–500 words)

Optimal Prompt Template (Level 2+)

Write a short story in {*language*} suitable for 5–7-year-old children.
Use simple, easy-to-understand words and limit the story to 3–4 short paragraphs (≈ 350–500 words).
The story should feature a clear beginning, middle, and end.
Incorporate the verb "{*verb*}", the noun "{*noun*}", and the adjective "{*adjective*}" naturally into the story.
Integrate the conclusion/tone "{*feature*}" through actions and outcomes without stating the tone explicitly.
Remember to keep the language age-appropriate.
Return the output as a JSON dictionary: {*"story": "your_generated_story"*}

Figure 4: **Optimal Prompt Template** used for the story-generation experiments in all three languages.

This template's effectiveness stems from several critical elements:

1. It specifies a clear target audience and language
2. It provides explicit structural guidance (3-4 paragraphs, clear beginning/middle/end)
3. It incorporates lexical constraints (verb, noun, adjective) to guide vocabulary usage
4. It requests thematic integration (feature/tone) through narrative rather than explicit statements
5. It maintains appropriate word count constraints (350-500 words)
6. It specifies the return format (JSON) for consistent processing

E.5 Implementation and Data Generation

For each language, 2 million unique prompts were generated and stored in JSON format. The implementation is available at `prompting/prompt-gen/create_prompts.py`[4].

Among state-of-the-art LLMs, GPT-4o-mini was selected due to its cost-effectiveness, multilingual capabilities, and robust API support. Story generation was parallelised using multiple concurrent API sessions. Specifically, a configuration of 4 sessions with 16 threads each was deployed on a 24 vCPU system, yielding a generation throughput of approximately 100 stories per minute. The implementation is provided in `prompting/request_helper.py`[1].

The final dataset comprises 2 million synthetic stories for each of Hindi, Bangla, and Marathi, generated via this scalable pipeline.

This end-to-end prompt-to-data generation framework addresses limitations of the foundational work, while ensuring both linguistic diversity and quality, thereby facilitating effective training of Small Language Models in these regional languages.

[4]Available in the project repository

F Analysis of Synthetic Training Data: Linguistic Diversity and Evaluation Metric Performance

Implementation for this section can be found at `analysis/` in our repository.

F.1 The Zero-ROUGE Phenomenon in Cross-Lingual Evaluation

Our experiments revealed a striking phenomenon when applying a traditional n-gram-based evaluation metric like ROUGE (Lin, 2004) to non-English text generation.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics designed to evaluate automatic summarisation and machine translation by comparing generated text to reference texts.

ROUGE-1 and ROUGE-2 measure the overlap of unigrams (single words) and bigrams (word pairs), respectively, between the candidate and reference texts, while ROUGE-L uses the longest common subsequence to assess sentence-level structural similarity.

As done in the foundational Tinystories paper, we wanted to utilise ROUGE to analyse the diversity/quality of the synthetically generated training dataset, to ensure our SLMs generate unique stories and not just memorise the training data.

Although evaluating text generation quality for English has established benchmarks, we observe significant challenges when applying the same metrics to, e.g., LLM-generated Bengali training stories from the TinyStories-Regional dataset.

F.1.1 Contrasting ROUGE Performance Between Languages

When applied to the English TinyStories dataset (Eldan and Li, 2023b), ROUGE metrics provided nuanced scores reflecting different degrees of lexical overlap:

Metric	Average F1
ROUGE-1	0.2916
ROUGE-2	0.0553
ROUGE-L	0.1700

Table 19: Average ROUGE F1 Scores (English)

Individual story scores exhibited a normal distribution of values, matching the reports from the Tinystories paper:

story_idx	rouge1_f1	rouge2_f1	rougeL_f1
0	0.272727	0.054054	0.124579
1	0.258503	0.006849	0.102041
2	0.375000	0.094488	0.218750
...	...	...	...
9	0.266160	0.061303	0.152091

Table 20: English TinyStories ROUGE scores sample

However, when the same methodology was applied to the entire Bangla TinyStories dataset (`TinyStories-Regional/beng-generated_4o-mini_2M`), ROUGE uniformly produced zero values:

Metric	Average F1
ROUGE-1	0.0000
ROUGE-2	0.0000
ROUGE-L	0.0000

Table 21: Average ROUGE F1 Scores (Bangla)

F.1.2 Contextual Analysis of Evaluation Metric Performance

To better understand this phenomenon, we conducted a comprehensive comparative analysis using other evaluation metrics. BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is an algorithm for evaluating machine translation quality based on n-gram precision, measuring how many generated phrases match reference translations. BLEU scores range from 0 to 1, with higher values indicating closer alignment to human references, though the metric tends to favour shorter texts and often fails to capture semantic equivalence.

BERTScore (Zhang et al., 2020) leverages contextual embeddings from pre-trained language models to compute similarity between generated and reference texts at a semantic level rather than exact word matches. This approach allows BERTScore to recognise paraphrases and synonyms as similar, making it more robust for evaluating text generation quality in morphologically rich languages where lexical variation is common.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) evaluates translation quality by calculating precision and recall weighted by importance, while also accounting for word order, stemming, and synonymy. METEOR typically correlates better with human judgments than BLEU by considering linguistic elements beyond n-gram matching, making it particularly useful for evaluating text in languages with flexible word order and rich morphology.

BLEU Score Analysis BLEU scores for Bangla stories exhibited considerable variation yet remained consistently low. The mean BLEU score was 0.078 ($\sigma = 0.126$), with values ranging from 0.003 to 0.421. Story index 5 demonstrated a notably higher BLEU score (0.421), suggesting some lexical alignment with its reference. The generally low BLEU scores corroborated our ROUGE findings, confirming significant lexical divergence (Fig. 6).

BERTScore Analysis In stark contrast to lexical metrics, BERTScore values were remarkably high across all Bangla story pairs. The mean BERTScore was 0.967 ($\sigma = 0.012$), with scores ranging from 0.944 to 0.982. This dramatic difference between BLEU and BERTScore revealed a fundamental characteristic of the generated stories: while they utilize different vocabulary and phrasing from the references, they maintain high semantic fidelity (Fig. 6).

METEOR Score Analysis METEOR scores occupied a middle ground between BLEU and BERTScore, with a mean of 0.153 ($\sigma = 0.046$) and range of 0.071 to 0.231. For the sample in Fig. 7, story index 2 achieved the highest METEOR score (0.231), while story index 9 received the lowest (0.071). The intermediate nature of METEOR scores reflects its design as a balanced metric that considers both lexical and semantic similarities.

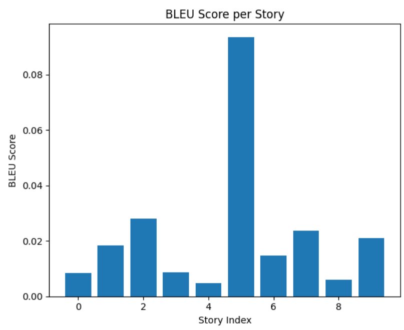

Figure 5: BLEU for 10 random Bangla stories.

The moderate correlation between BLEU and METEOR ($r = 0.63$) suggests that despite METEOR's consideration of synonymy, it still maintains sensitivity to lexical overlap. The weaker correlation between BLEU and BERTScore ($r = 0.29$) confirms that these metrics capture fundamentally different aspects of text similarity. Qualitatively

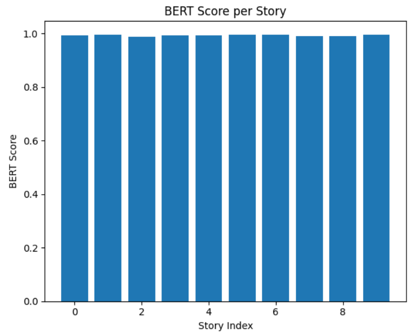

Figure 6: BERT for 10 random Bangla stories.

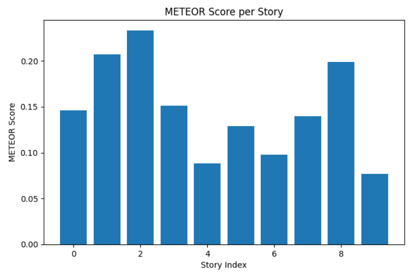

Figure 7: METOR for 10 random Bangla stories.

similar observations hold true for randomly sampled synthetic training data in Hindi (Fig. 8) and Marathi (Fig. 9).

F.2 Linguistic Factors Contributing to the Zero-ROUGE Phenomenon

The zero-ROUGE phenomenon observed in Bangla text evaluation can be attributed to several linguistic factors:

1. **Morphological Richness:** Bangla possesses a complex morphological structure with numerous inflectional and derivational forms, increasing the likelihood of lexical variation even when expressing identical concepts.
2. **Word Formation Patterns:** The agglutinative tendencies in Bangla create fewer opportunities for exact n-gram matches compared to English.
3. **Syntactic Flexibility:** Bangla permits greater variation in word order while preserving meaning, reducing the likelihood of matching n-grams even in semantically equivalent sentences.
4. **Training Methodologies:** Modern language models with multiple decoding paths may naturally produce diverse lexical realisations of similar semantic content, especially when the target language permits such variation.

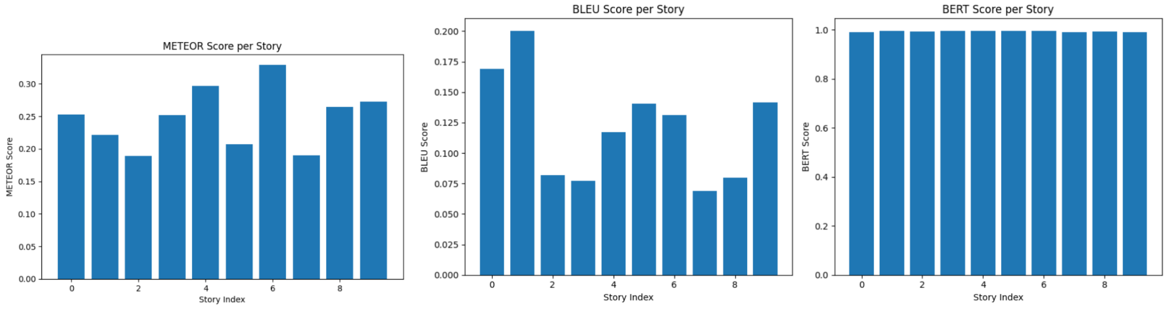

Figure 8: BLEU, BERT & METOR Scores for 10 random Hindi stories.

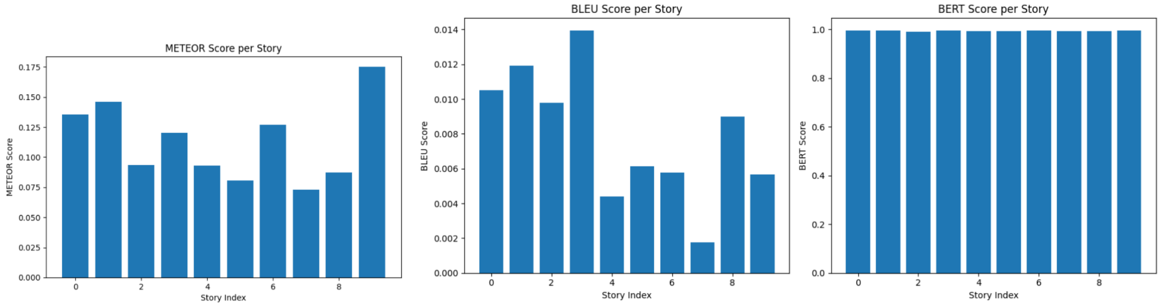

Figure 9: BLEU, BERT & METOR Scores for 10 random Marathi stories.

Metric Pair	Correlation Coeff.
BLEU-BERTScore	0.29
BLEU-METEOR	0.63
BERTScore-METEOR	0.51

Table 22: Pearson Correlation Coefficients Between Metrics for Bangla Stories

This finding represents an extreme manifestation of the limitations of lexical metrics, where the absence of exact n-gram overlap, as evidenced by zero ROUGE scores, suggests that text generation systems employ sophisticated paraphrasing mechanisms while maintaining semantic coherence.

F.3 Implications for Multi-Lingual Text Generation Evaluation

Our analysis suggests that robust evaluation of text generation requires a multi-metric, language-aware approach. Based on our findings, we propose:

F.3.1 Language-Specific Considerations

1. **Metric Selection:** Researchers must carefully select evaluation metrics appropriate to the target language, considering morphological complexity and typical paraphrasing patterns.
2. **Benchmark Calibration:** Distinct performance benchmarks should be established for each language rather than applying universal thresholds derived from English.
3. **Reference Design:** Evaluation datasets for morphologically rich languages should include multiple reference texts to better capture acceptable lexical variation.

F.3.2 Multi-Dimensional Evaluation Framework

For a comprehensive assessment of generated text quality across languages, we recommend an integrated approach:

1. **Semantic Fidelity Assessment:** Using embedding-based metrics like BERTScore with language-specific models to verify preservation of core meaning.
2. **Structural Evaluation:** Employing METEOR with language-appropriate resources for stemming and synonymy to assess whether narrative structure and word order are maintained within language-specific constraints.
3. **Lexical Diversity Measurement:** Calculating type-token ratios or using metrics like MTLD (McCarthy and Jarvis, 2010) to quantify lexical richness relative to language norms.

Sample Story:

একদিন, ছোট্ট রাহুল তার বন্ধুদের নিয়ে একটি ঐশ্বর্যময় ঝর্ণার কাছে গেল।
ঝর্ণার পানি ঝরঝরে এবং সাদা ফেনায় ভরা ছিল...

English Translation:

One day, little Rahul went to a glorious waterfall with his friends.

The water of the waterfall was sparkling and full of white foam...

Figure 10: A sample Bangla story from the synthetic dataset

Reference Story:

একদিন, রৌদ্রজ্জ্বল সকালে, ছোট্ট রিমি তার খাতা নিয়ে বাগানে বসেছিল।
সে খাতায় ছবি আঁকছিল। রিমির খুব ভাল...

English Translation:

One day, in the bright morning, little Rimi sat in the garden with her notebook.

She was drawing pictures. Rimi was very good...

Figure 11: A story for reference from the same dataset

4. **Reference-Free Quality Assessment:** Incorporating fluency and coherence metrics calibrated to the specific language being evaluated.

F.4 Case Study: Qualitative Analysis of Bangla Story Pairs

To illustrate the disconnect between lexical overlap and semantic similarity, we present a representative Bangla story pair from our dataset, along with the English translations in Figs. 10 & Fig. 11 (next page):

Despite sharing the theme of a child's experience outdoors, these stories use entirely different vocabulary, characters, and settings.

ROUGE metrics registered zero overlap, yet BERTScore identified high semantic similarity (0.961), recognising the shared narrative elements and emotional tone.

F.5 Final comments

Our discovery of the zero-ROUGE phenomenon highlights the need for better evaluation frameworks for non-English languages, particularly as text generation systems prioritise semantic preservation over lexical copying.

Analysis across Hindi, Bangla, and Marathi reveals consistent patterns:

- BLEU scores remain low (<0.2)
- BERTScore values approach near-perfect (>0.95)
- METEOR scores provide a middle ground (0.07-0.33)

This contrast demonstrates how traditional lexical metrics fail to capture semantic equivalence in morphologically rich Indian languages. The pattern confirms our generation approach produces semantically coherent content with lexical diversity, rather than relying on exact phrase repetition.

Future research directions should include:

- Developing specialized metrics balancing plot preservation with stylistic variation
- Establishing multilingual benchmark datasets with multiple reference texts
- Investigating human-metric correlations for generative tasks
- Exploring reference-free evaluation approaches

G Statistical Analysis of Inference Results

The implementation for this section is available at `results/compare_models_statistics.py` in our repository.

G.1 Key Limitation

Although identical procedures are employed to prepare data and train the Hindi, Marathi, and Bangla models, utilising GPT-4o evaluations (and their resulting statistics) to compare performance across these languages remains unreliable. This is due to the unknown relative "proficiency" of GPT-4o in each language, as well as potential disparities in the tokenisation capabilities of the Sarvam tokeniser across the three language scripts. Consequently, cross-lingual comparisons introduce an additional layer of noise and external bias that must be acknowledged.

In contrast, GPT–4o–based evaluations can be applied with confidence for intra-lingual comparisons, since within-language assessments are unaffected by these inter-lingual biases.

G.2 Distribution of Evaluation Metrics

We conducted a comprehensive statistical analysis of 3,000 stories generated by our 54M parameter Small Language Model (SLM) for each of the three target languages: Hindi, Bangla, and Marathi, using the Sarvam tokeniser.

This analysis provides deeper insights into the performance characteristics of our models across various evaluation dimensions.

G.3 Distributional Characteristics

The evaluation scores for all three languages exhibit distinct distributional patterns that reveal important aspects of model behaviour

In Table. 23, we summarise key statistical properties observed across languages and metrics.

G.4 Performance Patterns

Our analysis reveals several significant cross-linguistic patterns that provide insights into both model behaviour and inherent language characteristics:

1. **Hierarchical Emergence of Capabilities**: Across all three languages, we observe a consistent hierarchy in performance metrics, with grammar consistently achieving the highest scores (Hindi: 8.91, Marathi: 9.14, Bangla: 8.82), followed by fluency (Hindi: 8.55, Marathi: 8.11, Bangla: 8.85), completeness (Hindi: 7.78, Marathi: 8.127, Bangla: 7.64), and context awareness (Hindi: 7.73, Marathi: 7.15, Bangla: 7.51). This pattern aligns with the developmental progression observed in the foundational TinyStories research, suggesting that grammatical competence emerges earlier than contextual understanding, regardless of language.

2. **Bimodal Distribution of Context Scores**: Violin Plots (Figs. 12, 13 & 14) reveal a distinctive bimodal distribution for context awareness scores across all three languages, with concentration of scores around the 7 and 8-9 ranges. This bimodality suggests that stories tend to either achieve strong contextual coherence or struggle with maintaining context throughout the narrative, with relatively few stories falling in the intermediate range. This pattern is evident across all three languages but varies in intensity.

3. **Consistency in Grammar Scores**: Grammar scores exhibit the lowest standard deviation across all languages (Hindi: 0.34, Marathi: 0.50, Bangla: 0.42), indicating that once basic grammatical competence is achieved, it remains relatively stable across generated stories. The narrow distribution of grammar scores visible in the violin plots demonstrates the models' tendency to consistently produce grammatically correct text.

4. **Context–Is–Key**: Limitations (Sec.G.1) render direct, cross-lingual comparison of most individual ratings (Tables. 7, 8, and 9) unreliable. In contrast, context awareness evaluation—measuring fidelity to the original prompt, a more robust and less biased metric, since it focuses on each model's ability to maintain relevance to the original prompt rather than on absolute performance levels, provides a more dependable cross-lingual comparison metric. Our manual review shows that models with lower context scores, despite high standalone ratings, stray from the source and yield unusable outputs. Consequently, context is the paramount evaluation criterion, in which the Hindi 54 M-parameter model excels. This finding echoes our discussion of Rényi entropy: languages with higher entropy present greater coherence challenges, highlighting the

Language	Metric	Mean	Median	Std Dev
Hindi	Context Awareness	7.73	8.00	1.01
	Completeness	7.78	8.00	0.86
	Grammar	8.91	9.00	0.34
	Fluency	8.55	9.00	0.56
	Creativity	7.81	8.00	0.58
Marathi	Context Awareness	7.15	8.00	1.18
	Completeness	8.12	7.00	0.87
	Grammar	9.14	9.00	0.50
	Fluency	8.85	8.00	0.64
	Creativity	7.90	8.00	0.69
Bangla	Context Awareness	7.51	8.00	1.11
	Completeness	7.64	7.00	0.85
	Grammar	8.82	9.00	0.42
	Fluency	8.42	8.00	0.59
	Creativity	7.69	8.00	0.59

Table 23: Statistical Properties of Evaluation Metrics Across Languages for 54M models

Hindi model's strength.

G.5 Relationship Between Model Architecture and Evaluation Metrics

To understand how different architectural choices affect specific linguistic capabilities, we conducted correlation analyses between model parameters and evaluation metrics.

G.5.1 Parameter Efficiency Across Languages

Tables. 7, 8, and 9 in the main text illustrate the relationship between model parameter count and evaluation metrics for each language. Several key observations emerge:

1. **Divergent Scaling Patterns**: While all languages benefit from increased model size, the marginal improvements from scaling differ significantly. The comparable performance of similarly-sized models (54M parameters) across the three languages suggests that architectural scaling properties may be relatively consistent, though the absolute performance levels differ. Giving priority to context awareness, Hindi demonstrates the strongest performance at this parameter range, followed by Bangla and then Marathi.

2. **Optimal Parameter Allocation**: The inflection point in the performance-parameter curve occurs consistently around 54M (E: 512, L: 6) parameters across all three languages, with minimal improvements beyond this threshold. The 157M (E: 1024, L: 7) performs similarly to the optimal 54M model for all 3 languages, rendering it unoptimal owing to its high VRAM Usage and Training Time.

3. **Parameter Elasticity by Metric**: Different evaluation metrics show varying sensitivity to parameter scaling. Grammar scores demonstrate the lowest elasticity (average 12% improvement from 4.46M to 157M parameters across languages), while context awareness shows the highest (average 33% improvement). This supports our hypothesis regarding the hierarchical emergence of capabilities, with grammatical competence requiring the least model capacity and contextual understanding requiring the most.

G.6 Correlation Analysis Between Metrics

Hindi The evaluation metrics in Hindi short stories demonstrated significant intercorrelations. The strongest association is observed between creativity and overall quality assessment ($r = 0.73$, $t(2998) = 58.48$, $p < .001$), indicating that creative elements substantially influenced holistic quality perceptions. Grammar and overall quality demonstrated a robust positive relationship ($r = 0.69$, $t(2998) = 52.20$, $p < .001$), suggesting grammatical accuracy significantly contributed to quality judgments. Notably, completeness and fluency exhibited a strong correlation ($r = 0.72$, $t(2998) = 56.81$, $p < .001$), indicating narrative completeness typically accompanied smooth readability. The weakest relationship was identified between context awareness and completeness (r = 0.35, $t(2998)$ = 20.46, $p < .001$), suggesting these constructs captured distinct dimensions of narrative quality.

Bangla Analysis of Bangla short stories revealed similar correlation patterns, with creativity and overall quality showing the highest correlation coefficient (r = 0.80, $t(2998)$ = 73.01, $p < .001$). This suggests that creative expression was the predominant factor in quality assessment. Grammar and overall quality maintained a strong positive relationship (r = 0.71, $t(2998)$ = 55.20, $p < .001$), highlighting the importance of grammatical precision. Completeness and fluency demonstrated substantial correlation (r = 0.77, $t(2998)$ = 66.08, $p < .001$), reinforcing the connection between narrative coherence and reading experience observed across languages. Context awareness and completeness displayed a moderate correlation (r = 0.39, $t(2998)$ = 23.19, $p < .001$), indicating these metrics evaluated partially distinct aspects of narrative construction.

Marathi In contrast to Hindi and Bangla, Marathi short stories exhibited their strongest correlation between context awareness and grammar (r = 0.78, $t(2998)$ = 68.25, $p < .001$), suggesting a language-specific relationship between contextual appropriateness and grammatical structure. Creativity and overall quality maintained equivalent correlation strength (r = 0.78, $t(2998)$ = 68.25, $p < .001$), consistent with patterns observed in the other languages. Completeness and fluency correlation remained robust (r = 0.77, $t(2998)$ = 66.08, $p < .001$), indicating a consistent relationship across all three languages. The weakest association is observed between completeness and creativity (r = 0.49, $t(2998)$ = 30.78, $p < .001$), suggesting these dimensions function more independently in Marathi narratives compared to Hindi and Bangla.

These findings reveal both cross-linguistic patterns and language-specific relationships between evaluation metrics, with implications for understanding quality assessment in Indic-language short stories. Given the large sample size (n = 3000 per language), all correlations were statistically significant at $p < .001$, with the critical value for significance at this level being r = 0.060.

G.7 Comparative Analysis of Score Distributions

G.7.1 Distribution Variation Analysis

We examine the standard deviations across metrics and languages in Table 24, providing insight into the consistency of model performance.

Notable patterns include:

1. **Consistent Hierarchy of Variability**: Across all languages, Context Awareness shows the highest standard deviation (average: 1.10), indicating greater variability in the model's ability to maintain contextual coherence. Grammar consistently shows the lowest standard deviation (average: 0.42), suggesting that grammatical competence is more uniformly achieved once the model reaches sufficient capacity. On the other hand, context awareness is difficult to learn, and supported by Items.3, 4, it can be deemed as the most important evaluation criterion for the models.

2. **Language-Specific Consistency Patterns**: Marathi shows higher standard deviations across all metrics (average: 0.76) compared to Hindi (0.65) and Bangla (0.69), suggesting greater variability in performance. This is particularly evident in context awareness (Marathi: 1.18 vs. Hindi: 1.01) and overall scores (Marathi: 0.67 vs. Hindi: 0.52).

3. **Form vs. Content Metrics**: Metrics related to linguistic form (grammar, fluency) consistently show lower standard deviations (0.42, 0.60) than those related to content (context, completeness, creativity) (1.10, 0.86, 0.62). This pattern suggests that form-related capabilities may develop more uniformly and easily compared to content-related capabilities.

G.7.2 Consistency-Performance Relationship

Examining the relationship between metric means and standard deviations reveals important patterns:

1. **Inverse Relationship**: Across all languages, we observe an inverse relationship between mean scores and standard deviations. Metrics with higher means (like grammar) tend to have lower standard deviations, while metrics with lower means (like context awareness) show higher standard deviations. This pattern suggests that as performance on a particular aspect improves, consistency also increases.

Metric	Hindi	Marathi	Bangla	Average
Context Awareness	1.01	1.18	1.11	1.10
Completeness	0.86	0.87	0.85	0.86
Grammar	0.34	0.50	0.42	0.42
Fluency	0.56	0.64	0.59	0.60
Creativity	0.58	0.69	0.59	0.62
Overall	0.52	0.67	0.57	0.59
Average	0.65	0.76	0.69	0.70

Table 24: Standard Deviation Comparison Across Metrics and Languages

2. **Language-Specific Consistency**: All three languages show a moderate to strong negative correlation between means and standard deviations, with Marathi having the strongest inverse relationship (r = - 0.77), while both Bangla and Hindi show identical correlations (r = - 0.70). This negative correlation indicates that metrics with higher mean scores tend to have lower variability across all three languages, suggesting more consistent performance in areas where the models score higher.

3. **Metric-Specific Patterns**: Grammar shows both the highest means and lowest standard deviations across all languages, suggesting that grammatical competence represents a "foundational" capability that is both strong and consistent once achieved. Context awareness, by contrast, shows lower means and higher standard deviations, indicating, in accordance with Items.3, 4, & 1, it may represent a more advanced capability that remains challenging even as models improve.

G.8 Performance Gap Analysis

To better understand the relative strengths and weaknesses of the models across different languages, we analyse the gaps between different evaluation metrics (Table 25)

Key patterns include:

1. **Consistent Gap Hierarchy**: Across all languages, the largest performance gap is between grammar and context awareness (average: 1.32), while the smallest gap among the major metric pairs is between context awareness and completeness (average: -0.11, with context scores actually lower than completeness in all languages). This consistent pattern suggests a universal hierarchy in how different linguistic capabilities develop in these models.

2. **Language-Specific Gap Patterns**: Marathi shows notably larger gaps between grammar and other metrics (In Table. 8 — Grammar−Context = 1.99, Grammar−Completeness=1.09) compared to Hindi and Bangla. This suggests that while Marathi models achieve reasonable grammar scores, they struggle more with contextual coherence and narrative completeness compared to models in other languages.

3. **Form-Content Divide**: The substantial gaps between form-related metrics (grammar, fluency) and content-related metrics (context, completeness, creativity) highlight the models' stronger capabilities in producing structurally correct text versus semantically coherent narratives. This divide is most pronounced in Marathi and least evident in Hindi.

4. **Grammar-Fluency Relationship**: The gap between grammar and fluency scores is significantly smaller (average: 0.46) than between grammar and other metrics, suggesting these capabilities may develop in tandem. This pattern holds across all three languages, though Marathi shows a larger grammar-fluency gap (0.61) compared to Hindi (0.36) and Bangla (0.40).

G.9 Statistical Significance Analysis

To assess the significance of observed cross-linguistic differences, we analyzed the overall performance scores across the three languages (Table. 26).

These results suggest that:

1. The performance differences between languages, with context playing a crucial role (Item. 3), appear meaningful, with Hindi outperforming both Bangla and Marathi, and

Metric Pair	Hindi	Marathi	Bangla	Average
Grammar - Context Awareness	1.18	1.47	1.31	1.32
Grammar - Completeness	1.13	1.31	1.18	1.21
Grammar - Creativity	1.10	1.17	1.13	1.13
Grammar - Fluency	0.36	0.61	0.40	0.46
Fluency - Context Awareness	0.82	0.86	0.91	0.86
Fluency - Completeness	0.77	0.70	0.78	0.75
Fluency - Creativity	0.74	0.56	0.73	0.68
Context - Completeness	-0.05	-0.16	-0.13	-0.11

Table 25: Performance Gaps Between Metrics (Difference in Mean Scores)

Language	Mean	Standard Deviation
Hindi	8.158	0.52
Marathi	8.236	0.67
Bangla	8.016	0.57

Table 26: Overall Performance Statistics by Language

Marathi (Item. 2) showing the lowest overall performance.

2. The standard deviations indicate different levels of consistency across languages, with Hindi showing the most consistent performance (SD = 0.52) and Marathi showing the greatest variability (SD = 0.67).

3. These performance differences align with the entropy analysis presented in the main paper (Table. 5), where Marathi exhibited higher Rényi entropy values (7.76) compared to Hindi (7.15) and Bangla (7.41), suggesting a potential relationship between tokenization complexity and generation performance.

G.10 Final comments

Our statistical analysis reveals complex relationships between tokenisation strategies, linguistic properties, and generation performance across Hindi, Marathi, and Bangla. The consistent hierarchy of capabilities (grammar > fluency > completeness > context) across all three languages suggests universal aspects of language model development, while significant cross-linguistic differences in absolute performance point to the importance of language-specific optimisation.

Across the three languages (Tables 7, 8, and 9), the 54 M-parameter model scores highest overall in Marathi (8.236). Yet its context-awareness score (7.154) trails Hindi's (7.734), even though the two languages exhibit similar overall performance (Marathi 8.236 vs. Hindi 8.158). Acknowledging the constraints of our GPT-based evaluation (Sec. G.1), we note that context awareness emerges as the least-biased, most discriminative, and hardest-to-learn metric (Items.4,3,1 & 3). Coupled with Marathi's persistent statistical (Item.2) and inference (Sec.B.2), these findings justify the ranking (ease of model to ”learn” a language) Hindi > Bangla > Marathi, which echoes the Rényi-entropy analysis in the main paper: languages with higher entropy pose greater challenges for coherent text generation.

These findings underscore the value of the Regional TinyStories framework as both a practical approach to developing efficient language models for Indian languages and as an analytical tool for understanding comparative linguistic complexity. Future work should focus on exploring the relationship between tokenisation strategies, morphological characteristics, and model performance to develop more comprehensive metrics for predicting language modelling difficulty across typologically diverse languages.

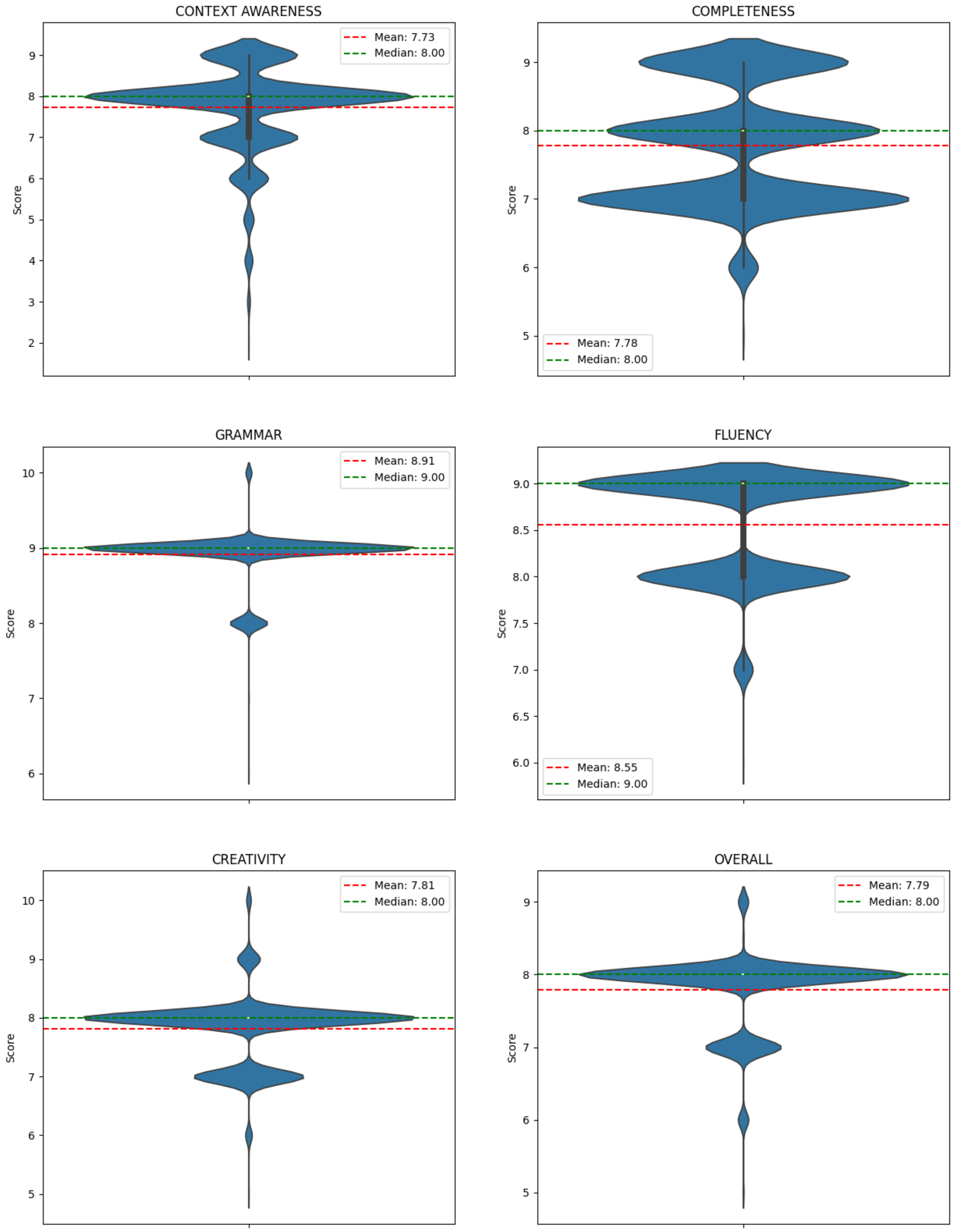

Figure 12: Violin plot depicting the inference score distribution for the 54M Hindi Model Inference (n = 3000).

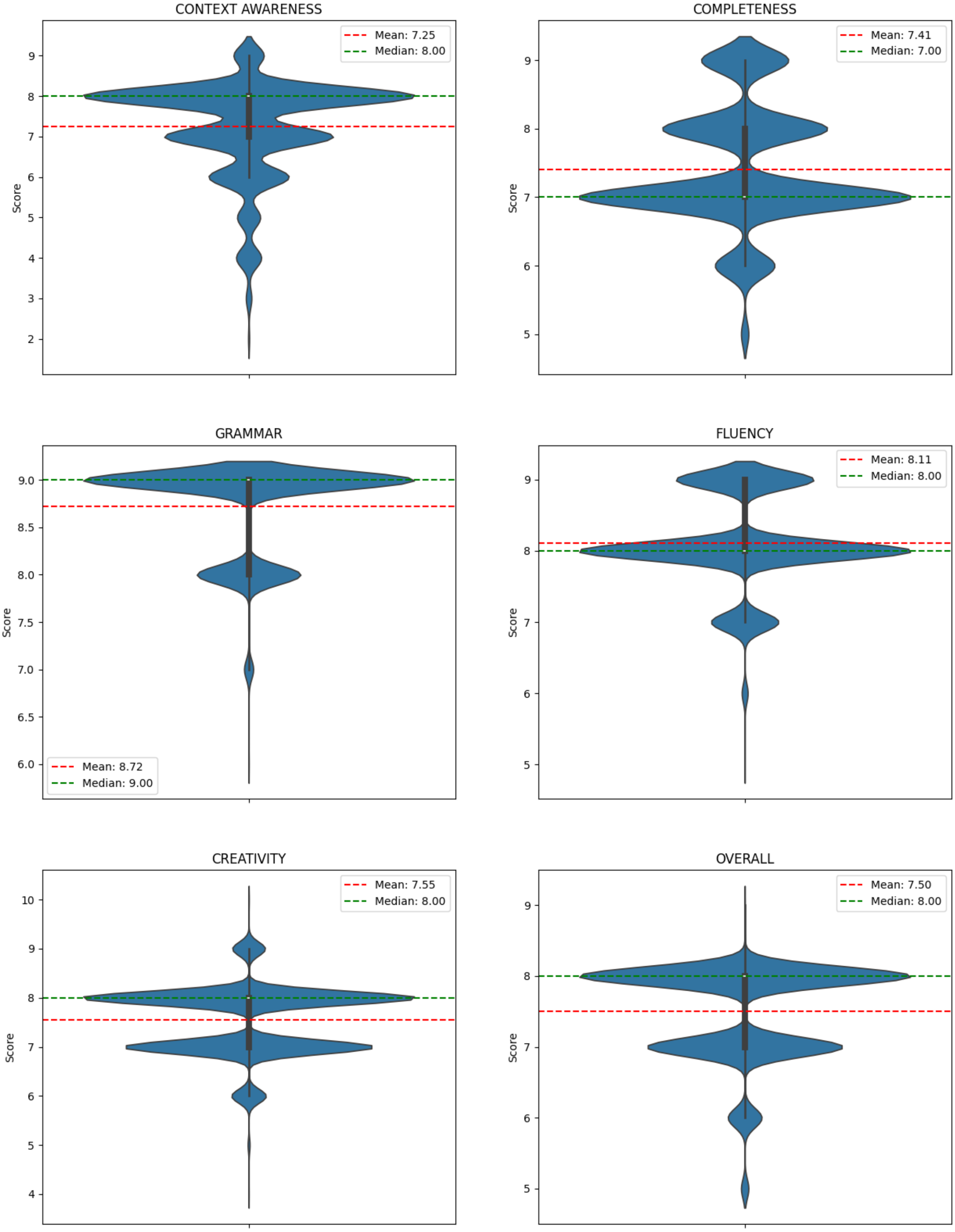

Figure 13: Violin plot depicting the inference score distribution for the 54M Marathi Model Inference (n = 3000).

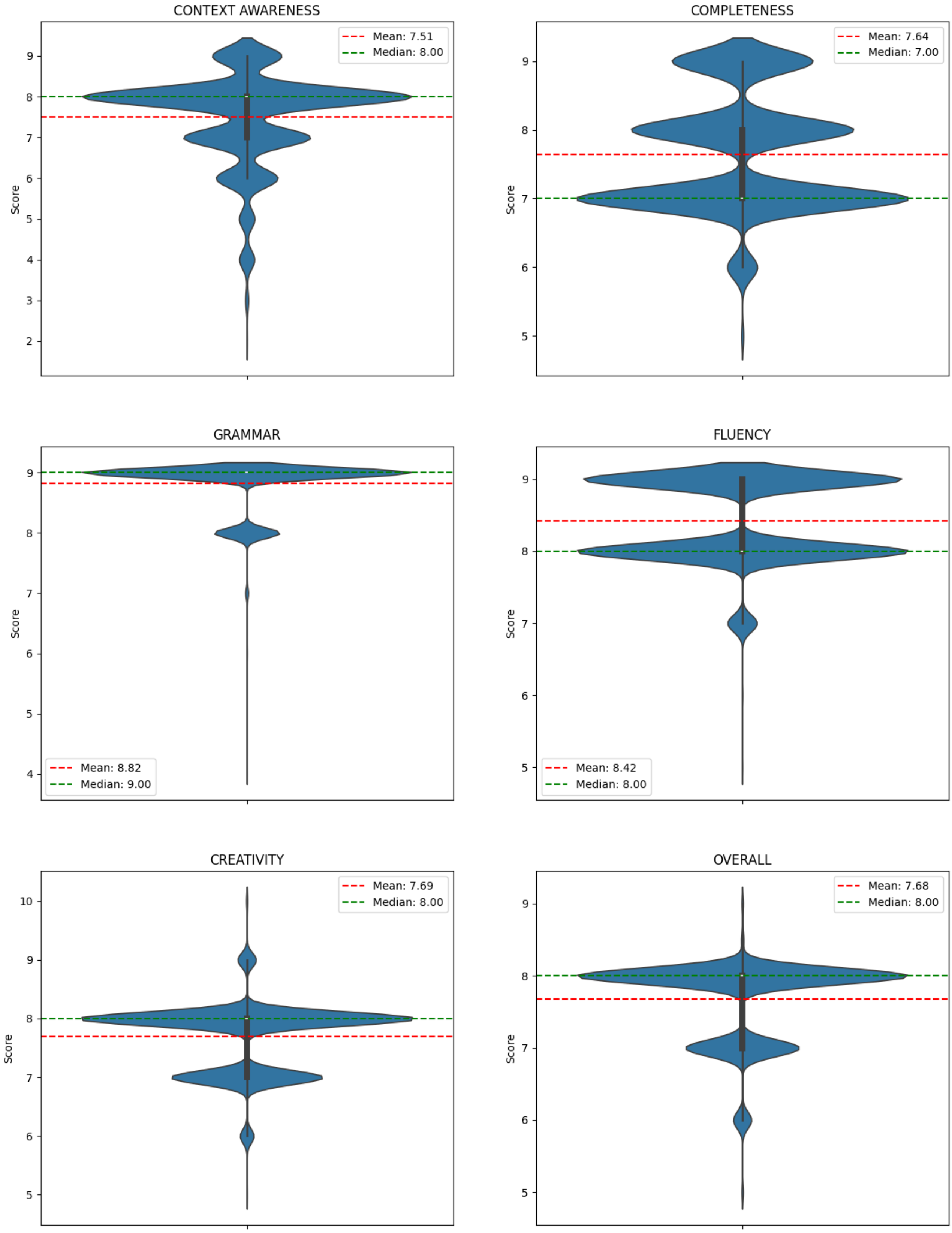

Figure 14: Violin plot depicting the inference score distribution for the 54M Bangla Model Inference (n = 3000).

H Comments on Tokenizers and Renyi Entropy

H.1 Tokenizer Analysis: *Sarvam* vs. *SUTRA* vs. *Tiktoken*

All three tokenizers (refer Sec. C.4 for details) rely on *sub-word* segmentation, yet they differ markedly in how much linguistic knowledge is baked into their vocabularies:

- **Sarvam-1** According to the model card, Sarvam uses a "vanilla" SentencePiece tokenizer trained on an Indic-centric corpus, resulting in a compact vocabulary of $\sim$68K tokens (4K reserved) and a token fertility of only 1.4–2.1 tokens per word for Devanagari and Eastern-Nagari scripts (Sarvam, 2024). Therefore, Hindi, Marathi, and Bangla inputs are encoded almost as economically as English, cutting sequence lengths and computation cost.

- **SUTRA** SUTRA is likewise SentencePiece-based, but it first learns sub-words from a multilingual corpus and then *merges* this inventory with the GPT-2 English vocabulary, yielding $\sim$ 256K tokens (Bendale et al., 2024). The larger table reduces splits on rare Indic morphemes at the price of higher memory. Empirically, its token fertility sits between Sarvam's and Tiktoken's.

- **Tiktoken (GPT-2)** OpenAI's tokenizer is a byte-level BPE trained almost entirely on English WebText and applied unchanged to every language (OpenAI, 2024). Prior work shows that such English-centric tokenizers can require 4–8$\times$ more tokens per word in Indic scripts, inflating latency and memory footprints (Petrov et al., 2023).

Observed impact In our experiments Sarvam attains the lowest evaluation loss among the two Indic tokenizers and the highest subjective fluency based on inference scores, suggesting its compact yet Indic-aware vocabulary best captures morphological boundaries. SUTRA follows closely—its larger inventory helps but incurs extra computation, whereas Tiktoken minimises perplexity numerically (due to possible over-segmentation) yet yields the weakest narrative quality, reinforcing claims that English-centric BPE introduces cross-language unfairness.

H.2 Renyi Entropy of Hindi, Marathi, and Bangla with SUTRA and Sarvam Tokenizers

When examining these metrics at different values of α (0.5, 1.0, 2.0) (Table. 27), we observe consistent patterns that illuminate different aspects of language complexity:

1. At $\alpha = 0.5$, which emphasizes rare tokens, Marathi shows the highest entropy (SUTRA: 7.2361, Sarvam: 7.5668), indicating a long tail of infrequent subwords. Notably, across all three languages, TikToken's entropy is substantially lower than that of SUTRA and Sarvam, revealing its coarser granularity and limited sensitivity to rare morphological forms.

2. At Shannon entropy ($\alpha = 1.0$), the relative ordering persists but differences moderate. Indic-specific tokenizers (Sarvam and SUTRA) maintain higher values—capturing richer morphological detail—whereas TikToken lags (e.g., Bangla: Sarvam 5.9807 vs. TikToken 2.8146).

3. At $\alpha = 2.0$, which weights common tokens, the gap narrows but remains: Indic tokenizers still reflect broader coverage of mid-frequency subword patterns, while TikToken's smaller vocabulary constrains its representational diversity.

Table 5 summarizes the Rényi entropy for Hindi, Bengali, and Marathi using TikToken (GPT-2), Sarvam-1, and SUTRA tokenizers. Vocabulary sizes differ notably: TikToken ($\sim$50K), Sarvam-1 ($\sim$68K), and SUTRA ($\sim$256K), influencing entropy values and reflecting distinct tokenization approaches.

TikToken shows the lowest entropy across all α values, indicating coarser tokenization with less morphological detail for regional Indian languages. Sarvam-1 balances vocabulary size and entropy, capturing meaningful words efficiently. SUTRA's largest vocabulary and highest entropy reflect a finer-grained, more complex distribution.

Marathi consistently exhibits the highest entropy, highlighting its morphological richness and correlating with observed modeling challenges.

Entropy trends at different α values further elucidate tokenizer characteristics: at $\alpha = 0.5$, Marathi's elevated values underscore its agglutinative complexity and long tail of rare subwords. At

Shannon entropy ($\alpha = 1.0$), persistent differences confirm that tokenizers with larger, more tailored vocabularies better capture Indic morphology.

According to Arnett and Bergen (2024), agglutinative languages have higher Rényi entropy compared to fusional languages. A study comparing Hindi, Marathi, and Bangla notes that Marathi's agglutinative structure creates more complex inflectional patterns, requiring distinct stemming strategies for information retrieval tasks. Bangla's simpler fusional morphology contrasts with Marathi's suffix-heavy word formation (Dolamic and Savoy, 2010).

Language	Tokenizer	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 2.5$
Hindi	Sarvam	6.9783	5.8996	4.6797	4.3566
	SUTRA	6.7723	6.0054	5.1897	4.9537
	TikToken	3.5927	3.0460	2.4131	2.2418
Marathi	Sarvam	7.5668	6.4795	4.9700	4.4069
	SUTRA	7.2361	6.4698	5.6302	5.1338
	TikToken	3.5780	3.0301	2.3414	2.1446
Bangla	Sarvam	6.8660	5.9807	4.7908	4.5083
	SUTRA	6.8095	6.0925	5.3475	5.3642
	TikToken	3.4969	2.8146	2.0116	1.7997

Table 27: Rényi entropy of Hindi, Marathi, and Bangla with Sarvam tokenizers

I Additional Benchmarking of SLM vs GPT-4 stories

I.1 COMET-DA and LaBSE

COMET-DA (Rei et al., 2022), originally developed for machine translation evaluation, has been repurposed in this work to compare generative outputs from language models across Hindi, Bangla, and Marathi under identical prompts. COMET-DA leverages pretrained cross-lingual encoders such as XLM-RoBERTa to produce contextualized representations and compute semantic similarity relative to reference texts or prompts. However, COMET-DA remains primarily tuned for translation adequacy and fluency, and under represents qualities central to narrative generation—such as discourse structure, stylistic nuance, and imaginative coherence. Additionally, encoder variance and normalization procedures calibrated to translation tasks may result in suboptimal scores even for semantically identical texts. Thus, in our case, we interpret COMET-DA scores as *a proxy for content fidelity* - i.e., how closely an SLM story preserves the semantic structure of a GPT4-generated story based on the same prompt.

To complement COMET-DA, we integrate LaBSE (Language-Agnostic BERT Sentence Embedding; (Feng et al., 2022)), a multilingual model trained on over 100 languages, including Hindi, Bangla, and Marathi. LaBSE enables cosine-based comparison of stories at the sentence or document level, providing a language-agnostic signal of *semantic similarity*. Despite its strengths, LaBSE—like COMET—does not account for creative structure or character development. While tools such as BLEURT (Sellam et al., 2020) and QUESTEval (Scialom et al., 2021) are designed to evaluate fluency and content preservation in English, they are not currently applicable to Indian languages due to their monolingual training data and English-specific QA pipelines. Neural embedding-based metrics correlate far better with human preferences than lexical n-gram metrics like BLEU/METEOR, particularly for open-ended generation in morphologically rich Indic languages.

As such, COMET-DA and LaBSE offer a viable alternative metric to LLM as a judge framework for evaluating generation quality in Indic languages due to robust multilingual support, tolerance to lexical variation and paraphrasing and efficient, reproducible scoring of large datasets. However, *they don't capture creativity, narrative coherence and other stylistic elements intrinsic to storytelling as they primarily rely on reference overlap*, which may penalise valid but divergent storylines. Nevertheless, to holistically assess story-level coherence, emotional resonance, and creativity, we recommend supplementing these metrics with human evaluation.

I.2 Final Conclusions

The following observations are made based on the results in Tables. 28 to 33:

- Hindi models show steady improvement in COMET-DA scores as parameters increase, with peak performance at 95M parameters before plateauing. Similarly, LaBSE cosine similarity improves with model size, reaching optimal performance at 54M and 95M checkpoints.

- Marathi models demonstrate an upward trend in COMET-DA scores but with notable fluctuations at larger sizes (95M and 157M). LaBSE similarity is highest at 54M model size.

- Bangla models consistently outperform both Hindi and Marathi, achieving higher COMET-DA scores that suggest better semantic alignment with GPT-4o outputs. Their LaBSE cosine similarity remains consistently high across all model sizes, indicating robust semantic similarity and stronger language modeling capabilities under the evaluated conditions.

- Across all languages, larger models generally achieve higher scores on both metrics, though performance plateaus or fluctuates at very large sizes (95M-157M). Bangla models consistently outperform the others, highlighting their superior semantic alignment with GPT-4o outputs, based on this metric.

  Overall, COMET-DA analysis points towards improving content fidelity with increasing model size, whereas LaBSE scores suggest comparable semantic similarity across all languages and models.

Hindi — COMET–DA			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.5649	0.5649	0.0546
4.65	0.5995	0.6012	0.0577
5.00	0.6185	0.6218	0.0527
9.00	0.6090	0.6132	0.0603
10.00	0.6311	0.6332	0.0578
19.00	0.6408	0.6445	0.0569
27.00	0.6508	0.6594	0.0680
41.00	0.6601	0.6638	0.0539
54.00	0.6556	0.6626	0.0634
66.00	0.6523	0.6624	0.0693
73.00	0.6472	0.6580	0.0727
95.00	0.6638	0.6686	0.0538
153.00	0.6591	0.6673	0.0632

Table 28: Hindi SLM checkpoints evaluated with COMET–DA (higher = closer to GPT-4o).

Marathi — COMET–DA			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.4860	0.4848	0.0458
4.65	0.5310	0.5311	0.0476
4.95	0.5410	0.5450	0.0585
41.16	0.5874	0.5943	0.0607
54.00	0.5793	0.5915	0.0723
73.00	0.5921	0.5980	0.0578
95.00	0.5599	0.5767	0.0866
157.00	0.5690	0.5752	0.0697

Table 29: Marathi SLM checkpoints evaluated with COMET–DA.

Bangla — COMET–DA			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.6849	0.6882	0.0533
4.65	0.7284	0.7335	0.0496
5.00	0.7139	0.7200	0.0542
41.00	0.7477	0.7590	0.0624
54.00	0.7559	0.7666	0.0590
73.00	0.7596	0.7677	0.0574
95.00	0.7373	0.7507	0.0702
157.00	0.7497	0.7630	0.0649

Table 30: Bangla SLM checkpoints evaluated with COMET–DA.

Hindi — LaBSE cosine			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.7404	0.7422	0.0456
4.65	0.7527	0.7562	0.0494
5.00	0.7470	0.7484	0.0472
9.00	0.7520	0.7532	0.0490
10.00	0.7579	0.7597	0.0497
19.00	0.7563	0.7586	0.0501
27.00	0.7492	0.7529	0.0554
41.00	0.7541	0.7558	0.0508
54.00	0.7682	0.7717	0.0541
66.00	0.7555	0.7594	0.0556
73.00	0.7551	0.7586	0.0586
95.00	0.7664	0.7698	0.0544
153.00	0.7617	0.7645	0.0545

Table 31: LaBSE similarity for Hindi checkpoints (ascending size).

Marathi — LaBSE cosine			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.7173	0.7198	0.0536
4.65	0.7617	0.7658	0.0544
4.95	0.7333	0.7362	0.0511
41.16	0.7476	0.7499	0.0516
54.00	0.7803	0.7849	0.0549
73.00	0.7450	0.7470	0.0539
95.00	0.7368	0.7402	0.0550
157.00	0.7519	0.7539	0.0553

Table 32: LaBSE similarity for Marathi checkpoints (ascending size).

Bangla — LaBSE cosine			
Params (M)	**Mean**	**Median**	**SD**
4.46	0.7711	0.7735	0.0459
4.65	0.7735	0.7766	0.0465
5.00	0.7765	0.7794	0.0480
41.00	0.7769	0.7810	0.0527
54.00	0.7776	0.7812	0.0543
73.00	0.7766	0.7799	0.0529
95.00	0.7674	0.7707	0.0548
157.00	0.7781	0.7823	0.0540

Table 33: LaBSE similarity for Bangla checkpoints (ascending size).