Regional-TinyStories: A Small Language Model Framework for Evaluating *Language Learning*, Tokenizers, and Datasets

Nirvan Patil[*,1,5], **Malhar Inamdar**[*,2,5], **Agnivo Gosai**[*,3,5], **Guruprasad Pathak**[4,5], **Anish Joshi**[4,5], **Anish Joshirao**[4,5], **Raj Dandekar**[5], **Rajat Dandekar**[5], **Sreedath Panat**[5]

[*]Equal Contribution, [1]BITS Pilani, [2]PICT Pune, [3]Independent Researcher, [4]PCCoE Pune, [5]Vizuara AI Labs

Correspondence: nirvan.ajit.patil@gmail.com, raj@vizuara.com

Links: Code (Github) & Datasets, Weights (HuggingFace)

Abstract

Small, resource-efficient language models are pivotal for extending high-quality text generation to low-resource and regional languages—the true frontier of linguistic equity in AI. Yet research largely prioritises massive English-centric systems, leaving regional-centric (low-resource) language modelling underexplored, particularly how tokenizer design, dataset diversity, and linguistic structure shape the inference of Small Language Models (SLMs) under realistic computational and data constraints. We present *Regional-TinyStories*, a lightweight framework that treats SLMs as cost-effective stand-ins for LLMs, enabling rapid, variable-wise inference-based analysis. Extending TinyStories to Hindi, Marathi, and Bangla, we release datasets of 2M synthetic and translated stories per language and train over 20 SLMs spanning 5–157M parameters. Using this framework, we (i) uncover contrasts between form-oriented (grammar, fluency) and content-oriented (context, completeness, creativity) metrics; (ii) chart language-specific learning dynamics; (iii) rank tokenizers, showing Indic-specific SARVAM-1 outperforming SUTRA and generic TIKTOKEN (GPT-2) across all metrics; and (iv) demonstrate that dataset semantic quality (translation vs. synthetic) strongly governs downstream generation. Validation through an LLM-as-Judge ensemble (GPT-4o, LLaMA-3.3-70B) and a 100+ participant human study confirms these trends while exposing systematic score inflation in automated evaluations. *Regional-TinyStories* offers a reproducible path to benchmark tokenizers, datasets, and SLM designs for scalable, context-faithful generation in low-resource settings.

1 Introduction

Research on language models (LMs) has largely emphasised scaling to multi-billion-parameter systems (Brown et al., 2020; Chowdhery et al., 2023), with performance improving as compute and data increase (Hoffmann et al., 2022). Yet the rising costs of training and inference—along with latency, memory, and energy constraints—are sharpening the focus on compact models that can be deployed under tight resource budgets. This need is particularly acute for low-resource and regional languages, where data scarcity and sovereignty concerns make smaller, efficient models especially appealing. In this work, we study "Small Language Models" (SLMs): low-million-parameter transformers trained under constrained data and compute regimes.

The TinyStories framework by Eldan and Li (2023b) demonstrates that English SLMs with fewer than 50M parameters can perform short-story inference when trained on small, curated datasets—echoing the modest linguistic exposure of children (fewer than 100M words by age 13) (Gilkerson et al., 2017). TinyStories shows that coherent text can emerge from compact architectures, with model width aiding knowledge retention and depth improving contextual understanding. Although today's "small" $\sim$5B-parameter models are still large in practice, ultra-compact, culturally aligned SLMs offer a promising path toward sovereign AI and digital inclusion. However, research on such models for regional languages remains limited (Boughorbel et al., 2024). We therefore extend TinyStories to regional languages to enable rapid, inference-driven comparative analyses of SLMs while amplifying under-represented regional languages in NLP and lowering the barriers to SLM development. Our aim is to understand how language choice, tokenizer design, and data curation shape downstream inference quality in regional languages, and to provide a lightweight evaluation setting supporting rapid design iteration.

Inference quality determines utility, yet isolating tokenization, dataset (language and translation-based) effects requires full-model ablations—

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2318–2367
December 20-24, 2025 ©2025 Association for Computational Linguistics

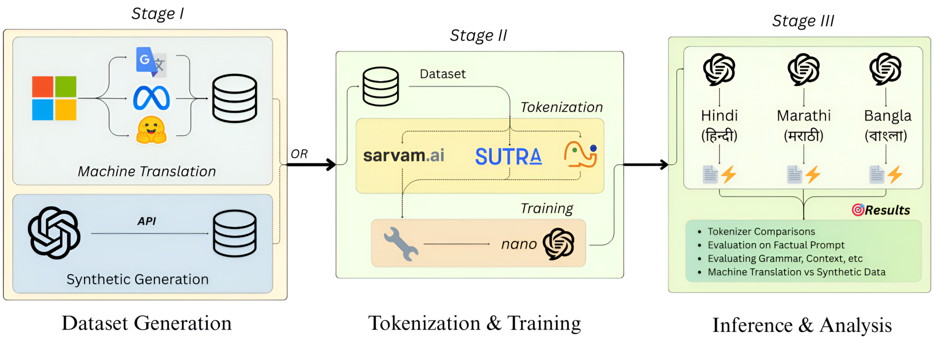


Figure 1: **Regional-TinyStories Framework**–Training distinct SLMs for variable-wise inference analysis.

training separate LLMs per variant—an impractically expensive path for systematic comparison. *We therefore introduce–Regional-TinyStories–a framework leveraging ultra-compact, TinyStories-based SLMs as time and cost-efficient proxies for LLMs* (see Fig. 1). By rapidly training one SLM per design factor, we enable direct, fine-grained assessment of how foundational design choices affect inference. These insights can then guide the resource-intensive development of larger models, mostly in line with established scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Hu et al., 2020; Tal et al., 2022). Our key contributions are:

- **Generalisation to regional languages:** We adapt the TinyStories paradigm to regional languages (here: Hindi, Marathi, Bangla), detailing data preparation and pre-training practices that yield high inference quality with small models.
- **Releases at scale:** We open-source the *Regional-TinyStories* codebase, over 10M synthetic and translated stories across the three languages, and weights for 35+ SLMs.
- **Scaling analysis:** We characterise inference performance as a function of model size and validate trends with statistical analysis, human evaluations, and qualitative examples.
- **Linguistic complexity:** We introduce an SLM inference–based measure of linguistic complexity (defined later) and show reduced inference quality for Marathi SLMs relative to Hindi under comparable settings.
- **Tokenizer comparisons:** We provide a fine-grained comparison across tokenizers, demonstrating that inference-quality evaluation offers actionable insight beyond traditional information-theoretic metrics, contrasting a legacy tokenizer (Tiktoken/GPT-2 (OpenAI, 2024)) with Indic-specialised tokenizers (Sarvam-1 (Sarvam, 2024), SUTRA-mlt-256-v2 (Bendale et al., 2024)).
- **Dataset comparisons:** We employ inference-quality evaluation as a proxy for dataset quality, showing that SLMs trained on synthetic data substantially outperform those trained on translated data and quantifying the adverse impact of translation-induced semantic loss on downstream inference.

2 Related Work

Indic NLP remains under-represented in mainstream NLP, with comparatively fewer curated, open resources. Recent efforts such as the IndicNLP Suite and benchmarks (Kakwani et al., 2020) and the monolingual corpora and models culminating in IndicBERT v2 (Doddapaneni et al., 2023) have begun to close this gap, while large parallel resources like *Samanantar* have energised MT and cross-lingual research (Ramesh et al., 2022). Our work complements these by providing story-style, generation-focused corpora (Hindi, Marathi, Bangla) and matched SLM releases that are especially accessible for early-stage students and researchers, enabling controlled, low-cost experimentation on inference behaviour.

SLMs gained momentum with BabyLM (Warstadt et al., 2023) and TinyStories, which showed small-scale data can elicit coherent generation from sub-50M models in English. We extend this paradigm to Indic languages and improve over TinyStories and other efficient-model lines through *inference-based* comparisons across languages, tokenizers, and datasets—prioritising downstream generation quality rather than proxy pretraining metrics.

Tokenization strongly shapes multilingual efficiency and quality; recent work also probes tokenizer adequacy via morphology-aware scores (e.g., MorphScore) to explain cross-language performance gaps (Arnett and Bergen, 2025). In contrast, we directly *rank tokenizers by their impact on downstream inference quality* in generative settings, providing application-grounded guidance for regional languages.

LLM-as-judge has emerged as a scalable alternative to human evaluation but exhibits known biases; MT-Bench/Chatbot Arena and subsequent analyses document both utility and vulnerabilities (Zheng et al., 2023a; Chen et al., 2024). We employ an ensemble of LLM judges and cross-validate our findings with a human evaluation involving over 100 participants, confirming trends while exposing systematic inflation, which aligns with the cautions reported in prior work.

3 Regional-TinyStories Datasets

3.1 Dataset Preparation

Synthetic Data

Our data generation pipeline, applied uniformly across Hindi, Marathi, and Bangla, builds on the TinyStories methodology while addressing several of its limitations. Specifically, we introduce (i) manual verification and full transparency of the prompt-generation vocabulary, (ii) a systematic ablation study to select an optimal prompt template, and (iii) a custom mechanism to prevent duplicate prompts—together improving reproducibility, data diversity, and downstream quality.

The matched process for each language begins with a rigorously curated and manually verified lexicon/vocabulary comprising ~300 of each noun, verb, and adjective, augmented with carefully chosen stylistic "features" (e.g., *playfully mysterious*, *colorfully gentle*) that guide tone and moral framing.[1] Using results from our ablation study (Fig. 2, left), we select the best-performing template–identified via measuring downstream GPT-4o story generation quality–and instantiate it by sampling items from the lexicon, yielding a corpus of 2 million prompts. Uniqueness is enforced via a custom duplicate-avoidance algorithm to ensure coverage and variety.[2] These prompts are then passed to GPT-4o-mini to synthesise the corresponding 2M stories for each language.

The Regional–TinyStories corpus, when tokenized using Sarvam-1, comprises ~658M tokens in Hindi, ~639M in Marathi, and ~618M in Bangla, with each story averaging ~320 tokens. Stories are aimed towards children aged 5–7 and are designed to cover the age–relevant stylistic "features" in our lexicon with near–uniform frequency, as confirmed by a manual evaluation of 1,000 stories per language.

Translated Data

Following prior work for translation to Indic languages (Doshi et al., 2024; Boughorbel et al., 2024), we translated the complete TinyStories dataset (Eldan and Li, 2023a) (~2M stories) from English to each of Hindi, Marathi and Bangla using NLLB-200-3B (Team et al., 2022) and Google Translate.

3.2 Dataset Evaluation

Synthetic Data

We evaluate the synthetic corpus along lexical diversity and semantic coherence. From random pairs of training stories, we compute pairwise lexical similarity using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), and semantic similarity using BERTScore (Zhang et al., 2020). We observe low lexical overlap (BLEU: $\bar{x} = 0.078$, METEOR: $\bar{x} = 0.153$), indicating high lexical diversity, alongside high semantic alignment (BERTScore: $\bar{x} = 0.967$), consistent with coherent, age-appropriate narratives—likely reflecting GPT-4o-mini's guided, theme-based generation for 5–7-year-old readers.

To test for memorisation, we compare a random training story against an SLM-generated story. The same pattern emerges: low BLEU and METEOR confirm that SLM outputs are not surface copies of the training set, while a high BERTScore shows that generated stories preserve age-appropriate semantics and thematic coherence relative to the corpus.

We further conduct a large-scale quality assessment with GPT-4o. Using the prompt template in Fig. 2 (right), GPT-4o rates 3,000 randomly sampled stories on Context, Creativity, Completeness, Fluency, and Grammar, yielding consistently strong scores ($\bar{x} \geq 8.5/10$) across all dimensions, in line with the GPT-4o-mini results reported for each language in Fig. 3. Finally, a manual review

[1] Available at `prompting/prompt_gen` in our Codebase.

[2] Details on the lexicon, ablation setup, and duplicate-avoidance algorithm are provided in Appendix E.

Optimal Prompt Template (Level 2+)

Write a short story in {*language*} suitable for 5–7-year-old children using simple, easy-to-understand words and limited to 3–4 short paragraphs (approx. 350–500 words) with a clear beginning, middle, and end; naturally incorporate the verb "{*verb*}", the noun "{*noun*}", and the adjective "{*adjective*}"; integrate the conclusion/tone "{*feature*}" through actions and outcomes without stating it explicitly; keep the language age-appropriate; and return the output as a JSON dictionary {"story":"your_generated_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*}". Only return a JSON dictionary in the following format: {"context awareness":, "completeness":, "creativity":, "fluency":, "grammar":}.

Figure 2: Optimal Prompt Template for **Story Generation** and **LLM-as-a-Judge Evaluation**.

of 300+ training stories across languages verifies that samples are high-quality, thematically appropriate, and free of harmful content.[3]

Translated Data

We assess the semantic fidelity of our translated TinyStories corpora using Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2022), whose multilingual encoder captures cross-lingual semantics beyond surface n-gram overlap. LaBSE is thus better suited for translation evaluation than lexical metrics (BLEU, METEOR) or general-purpose multilingual BERT variants, whose optimisation often favours high-resource languages. For each target language, we compute cosine similarity on 1,000 English–translation story pairs and derive bootstrap-based 95% confidence intervals.

LaBSE reveals statistically significant differences in fidelity across languages. Bangla yields the highest mean similarity (0.9002), followed by Marathi (0.8824) and Hindi (0.8793); non-overlapping confidence intervals confirm all pairwise gaps at $p < 0.05$. Distributional inspection shows heavy tails, and Shapiro–Wilk rejects normality, indicating heterogeneous quality and occasional translation failures (minima ≈ 0.75).[4]

4 Methodology and Setup

4.1 SLM Architecture and Training

We train SLMs from scratch in PyTorch using a nanoGPT-style pipeline (Andrej Karpathy, 2022), adapting an open-sourced and user-friendly codebase. Each model is a decoder-only Transformer with 8 attention heads and varying embedding dimensions (64, 512, 768) and depths (2, 6, 12), spanning parameter counts from 5M to 157M. All models are trained for 5,000 steps; we hold out 2.5% of a single 2M-story dataset for testing.[5]

[3]Details, analyses and results are provided in App. F.

[4]Details, results and plots are provided in Appendix G.

[5]Details on hyperparameter configurations, optimisers, loss curves, and more are provided in Appendix D.

4.2 Tokenization

We evaluate three tokenizers: two open-source, SentencePiece-based systems—Sarvam-1 (Sarvam, 2024) and SUTRA-mlt-256-v2 (Bendale et al., 2024)—and OpenAI's widely used *Tiktoken* baseline (OpenAI, 2024). Sarvam-1 targets ten major Indic languages plus English (Bangla `bn`, Gujarati `gu`, Hindi `hi`, Kannada `kn`, Malayalam `ml`, Marathi `mr`, Odia `or`, Punjabi `pa`, Tamil `ta`, Telugu `te`, and English) with a vocabulary of ~64k. By contrast, SUTRA-mlt-256-v2 covers 50+ languages—including key Indic languages such as Hindi, Gujarati, Bangla, and Tamil—with a ~250k vocabulary (details in Appendix D.4.2). Given its narrower linguistic scope and Indic-focused vocabulary, Sarvam-1 is consistently superior on our Indic evaluations—across efficiency and content-oriented metrics—also offering substantially faster tokenization (see Sec. 5.6). We therefore adopt Sarvam-1 as the default tokenizer for all subsequent experiments, unless otherwise specified.

4.3 Inference-based Evaluation

After training, each SLM generates three distinct continuations for each of 1,000 manually curated prompts (all held out from training),[6] yielding 3,000 stories per configuration.[7] "Distinct" denotes differences in wording and meaning; decoding uses a fixed temperature (0.80) within each run. The 1,000 prompts are identical across SLMs and are language-matched via translation for Hindi, Marathi, and Bangla.

We then adopt an LLM-as-judge protocol (Eldan and Li, 2023a; Zheng et al., 2023b): GPT-4o evaluates every story using the rubric in Fig. 2 (right), assigning 1–10 scores for Context, Completeness, Creativity, Fluency, and Grammar. The overall score is the mean of these five. For each SLM configuration, we report (see Fig. 3) micro-averaged results—simple averages over its 3,000

[6]Available at `training-inference/prompt-<lang>`

[7]Inference samples are provided at `results/`

stories for each dimension and overall—which serve as the basis for downstream comparisons.

5 Results & Discussions

5.1 Performance Saturation

The inflection point in the performance-parameter curve occurs consistently around 54M parameters across all three languages, with minimal improvements beyond this threshold (see Fig. 3). We identify the 54M-parameter models, trained using the Sarvam-1 tokenizer, as the optimal configuration, which we adopt as the default for subsequent analyses.

5.2 Emergence of Capabilities

We synthesise and graph the quantitative patterns in Fig. 3,[8] organising findings by "form-oriented" metrics (Grammar, Fluency) versus "content-oriented" metrics (Completeness, Creativity, Context). Form captures syntactic correctness and linguistic fluency; content reflects semantic adequacy and contextual reasoning. The trends below summarise cross-language, cross-size behaviour and the relative ease with which form stabilises compared to content, as well as how this gap evolves with model scale.

- *Metric Order:* Grammar ($\bar{x} \approx 9.0$) $>$ Fluency (≈ 8.6) $>$ Completeness (≈ 7.9) $>$ Context (≈ 7.5) across all languages and model sizes (see Fig. 3). *This mirrors TinyStories' findings, where grammatical skills emerge before contextual understanding across languages.*

- *Variability:* High means correlate with low dispersion (Grammar $\sigma \approx 0.42$ vs. Context $\sigma \approx 1.18$. *Performance and consistency develop in tandem; Context may represent an advanced capability that remains challenging despite model improvements.*

- *Consistent Hierarchy of Variability:* Grammar and Fluency remain high (>8) even in smaller 5M-parameter models, while Creativity, Completeness, and especially Context emerge only at larger scales. *Grammatical accuracy stabilises early, whereas context-sensitive reasoning remains more complex and an emergent trait.*

- *Parameter Elasticity:* Grammar improves 12% and Context 33% when scaling from 4.46M to 157M parameters. *This corroborates the previous point, grammatical competence requires the least capacity while contextual understanding requires the most.*

- *Form vs. Content:* Across model sizes, 'form metrics' (Grammar, Fluency) show tighter bands ($\sigma \leq 0.60$) than 'content metrics' (Completeness, Creativity, Context) ($\sigma \geq 0.62$). *Models reliably produce structurally correct text rather than semantically coherent narratives.*

- *Performance Gaps:* Largest gap Grammar–Context ($\Delta \approx 1.32$); least Grammar–Fluency ($\Delta \approx 0.46$) across all languages. *Metrics with lower gaps are likely to develop in tandem.*

Taken together, the micro-averaged results in Fig. 3 underscore that structural well-formedness (syntactic correctness and linguistic fluency) emerges sooner than context-sensitive reasoning and should be interpreted as a priority in inter-SLM and cross-language comparisons.[9]

5.3 LLM-as-Judge Ensemble

We use GPT-4o to both generate stories and to evaluate SLM inference, which can introduce model-family bias. To mitigate this, we report scores from an independent judge, LLaMA-3.3-70B (see Fig. 3, Tab. 6). The two judges receive *identical* evaluation rubric, prompts, and inputs, and LLaMA re-scores the *complete* evaluation set previously graded by GPT-4o.

While LLaMA yields uniformly lower absolute scores, it preserves the relative ordering and the central trends noted previously (Sec. 5.2). Consistent with the aim of avoiding single-family artefacts, we prioritise this agreement in trends over absolute magnitudes—the criterion that governs downstream, metric-based, *relative* comparisons.

5.4 Human Evaluations

Recent advances suggest that LLMs can act as competent proxy judges for open-ended language tasks: *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* reports human–LLM agreement reaching roughly 80% on pairwise preferences, indicating viability at scale (Zheng et al., 2023a). At the same time, systematic biases remain–Chen et al. (2024) documents judgment biases and vulnerability to prompt manipulations.

[8] See Appendix A for tables reporting numerical results.

[9] See App. H for additional details and statistical analyses.

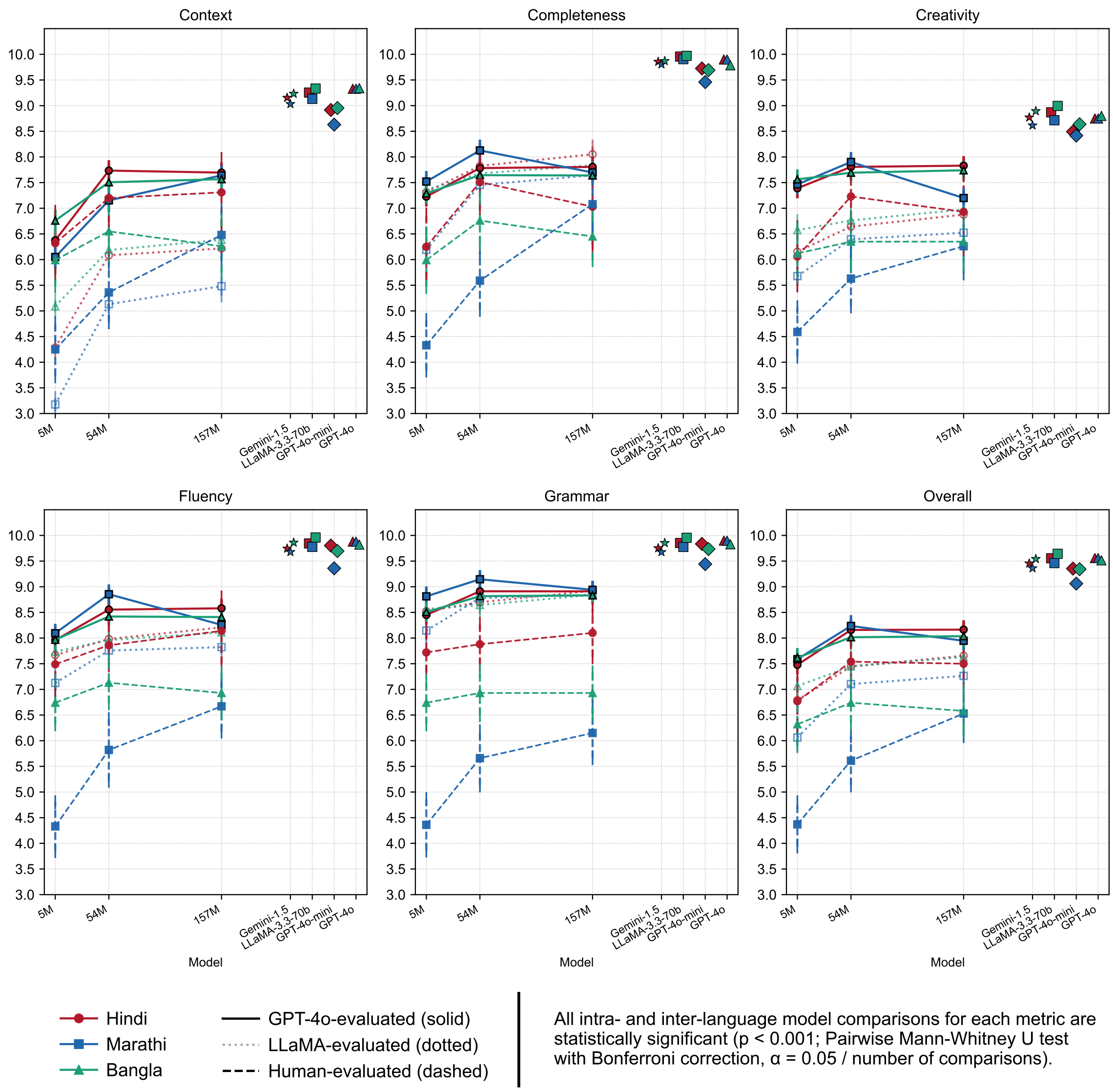


Figure 3: **Inference evaluation scores** across key-SLM sizes trained on one of Hindi, Marathi or Bangla *Regional-TinyStories* synthetic datasets. SOTA LLM results denote performance on stories generated from the same prompts as the SLMs, assessed by GPT-4o. *Complete numerical results are provided in Appendix A*

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 colourful 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 realised 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 1: **Hindi SLM generated Story** – *See Appendix B for additional SLM short-story inference examples across languages and model sizes.*

Complementing these findings, our own LLM-judge ensemble shows that, while LLaMA yields uniformly lower absolute scores than GPT-4o, it preserves relative orderings and the central trends noted earlier (Sec. 5.2). Consistent with the aim of avoiding single-family artefacts, we prioritise *trend replication* over absolute magnitudes—the criterion that governs downstream, metric-based, *relative* comparisons, conducting a dedicated human evaluation as a second pillar of verification.

We collected responses from 100+ fluent speakers who voluntarily participated in the study, collectively providing 650+ ratings via 8–10 minute Google Forms sessions (ratings reported in Fig. 3). Each rater applied the *same rubric text as used for the GPT-4o judge (Fig. 2 (right)*, scoring *Context*, *Creativity*, *Completeness*, *Fluency*, and *Grammar*, on 10-point Likert scales. To mitigate fatigue and inconsistency, we supplied concise metric definitions and brief calibration examples; attention checks were used to filter unreliable submissions following best-practices–van der Lee et al. (2019).

Human scores are lower in absolute value than GPT-4o's, but they replicate the central patterns captured by the LLM judges (Fig. 3): early mastery of form—*Grammar* and *Fluency*—and persistent difficulty on content—*Context* and *Completeness*. After Z-score normalisation, Kruskal–Wallis tests reveal significant differences across model sizes and languages ($p < 0.001$), substantiating the observed separations in the trend-oriented comparisons. Inter-rater reliability (Krippendorff's α) is *Strong* for Bangla (α=0.705), *Good* for Hindi (α=0.666), and lower for Marathi (α=0.332), indicating heterogeneous proficiency and motivating tighter screening and targeted guidance for future Marathi evaluations. In terms of normalised *Context* awareness, Hindi leads, followed by Bangla and then Marathi, corroborating the ordering surfaced by our framework.

In sum, the human study confirms *trend replication*—the key criterion for our downstream, relative analyses—while exposing LLM-as-a-Judge score inflation. Automated LLM-as-a-Judge evaluation thus serves as a reliable, high-throughput mechanism for comparative assessment for Indic short stories, with periodic human calibration remaining essential for anchoring scales and stress-testing judge robustness.[10]

[10]See App. C for tables reporting human ratings, deltas w.r.t. LLM-eval, Google Forms, etc.

5.5 SLM *Language-Learning* Analysis

We examine SLMs' language-learning by assessing whether models trained in a target language can generate stories that are *context-aware, complete, creative, fluent, and grammatically correct*. Cross-language comparisons are inherently delicate because LLM-as-a-Judge systems and tokenizers differ in maturity across languages; therefore, observed gaps are best interpreted as stemming from today's technological stack rather than intrinsic linguistic difficulty under controlled conditions.

A broad evaluation at 54M parameters reveals that Marathi-trained SLMs lag Hindi (and, to a lesser extent, Bangla). Two signals dominate. First, a pronounced *Context–score gap*: Grammar is relatively strong but diverges from Context and Completeness (Grammar–Context $\Delta \approx 1.47$; Grammar–Completeness $\Delta \approx 1.31$), indicating surface form is learned more readily than prompt-faithful content. Second, poorer *consistency*: mean–SD correlations are strongly negative across languages, but most extreme for Marathi ($r \approx -0.77$ vs. ≈ -0.70 for Hindi/Bangla), implying less stable instruction-following at a given quality level. Human raters validate this trend: the 54M Marathi SLM's contextual scores fall below its Hindi and Bangla counterparts. Although inter-rater agreement is weaker, the 95% CI mean lying entirely within the corresponding Hindi and Bangla intervals offsets this limitation (Fig. 3).

A story that violates its prompt is unusable; we foreground *Context Awareness*—adherence of the narrative's middle and end to the input prompt—as the primary yardstick. We measure it through three complementary probes. (i) *Short factual prompts* are especially informative: they probe context tracking, test logical comprehension beyond memorisation, and expose hallucination tendencies; here, Hindi clearly outperforms Marathi (Tab. 2). (ii) *Manual verification* of 100+ 54M-parameter Marathi SLM stories highlights frequent prompt drift, especially in the story's middle and end. (iii) *Large-scale evaluations*—LLM-as-a-Judge and trend-following human studies—converge on Marathi's weaker contextual adherence relative to Hindi and Bangla, most starkly at 54M parameters (Fig. 3).

Rapid, inference-based comparisons in the *Regional-TinyStories* framework thus yield a clear ranking by context-following: Hindi performs the strongest, Bangla is intermediate, and Marathi lags.

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 बिस्तर 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!

Table 2: Cross-lingual performance across key model sizes on **factual prompts** — *GPT-4 = 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
See Appendix B for additional examples across all configurations.

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 3: **Tokenizer performance** across languages for 54M parameter SLM
Sarvam Tokenizer — 54M (E=512, L=6) Model — Lighter is better.

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 4: **Translated vs. Synthetic** training data comparison for 54M parameter SLMs.
Sarvam Tokenizer — 54M (E=512, L=6) Model — Lighter is better.

The Bangla–Marathi gap narrows as model size increases from 54M to 157M, suggesting that capacity helps but does not close Marathi's deficit without language-specific data and modelling advances. Taken together, these results indicate that, under today's data and tooling, Marathi SLMs warrant targeted development.

5.6 Comparative Analyses of Tokenizers

Tokenization is typically assessed with information-theoretic tools; here we pair these with morphology-aware evaluation and then connect intrinsic signals to downstream behaviour. Rényi entropy H_α quantifies uncertainty in token distributions (with α modulating sensitivity to rare tokens) (Zouhar et al., 2023), while MorphScore measures alignment of token boundaries with linguistic morphemes (Arnett and Bergen, 2025). Applying *byte-premium scaling* to normalise corpora at the byte level (Arnett and Bergen, 2025), Sarvam yields more compact segmentations across languages (Hindi: 6.285, Marathi: 6.545, Bangla: 6.358), indicating efficient subword capture and compact tokenization, whereas SUTRA's higher entropies (Hindi: 7.153, Marathi: 7.762, Bangla: 7.414) reflect richer representation but increased complexity. On MorphScore, Sarvam matches or exceeds SUTRA in Hindi (0.728 vs. 0.727) and Bangla (0.319 vs. 0.300), and is comparable in Marathi (0.662 vs. 0.667). While these corpus-bound metrics surface key morphological patterns, they do not directly *quantify* inference differences.

To ground them in behaviour, *Regional-TinyStories* re-tokenises a fixed dataset with each tokenizer, trains matched SLMs, and compares inference. Results mirror the intrinsic picture: Sarvam-tokenised models generally edge SUTRA-cruically identifying the gains as concentrated on content-oriented metrics—Context, Completeness, Creativity (mean Sarvam-SUTRA difference of $+0.247$)—and smaller gains on form-oriented metrics—Fluency, Grammar (mean $+0.150$; Tab. 3). This accords with Sarvam's Indic specialisation, while despite SUTRA's broad multilingual design (50+ languages), it trails only modestly on Indic tasks. Compared to a non-Indic baseline, Indic-tokenized SLMs (Sarvam, SUTRA) outperform Tiktoken-trained ones across all metrics, despite Tiktoken's lower eval loss. The largest gains appear in *Context Awareness*, with the narrowest gap in Grammar. In sum, these behavioural results complement Rényi entropy and MorphScore, by empirically linking tokeniser design to end-task behaviour (inference), providing a concrete, actionable basis for tokenizer selection by highlighting the gains delivered by language-specialised tokenization being linked to Contextual Awareness.

5.7 Effect of Training on Translated vs Synthetically Generated Datasets

Translation remains susceptible to *semantic loss* (Cohn-Gordon and Goodman, 2019; Cao et al., 2020)—meaning drift, context distortion, and misaligned entailments—even with modern systems. Sentence-level embedding metrics (LaBSE) quantify this loss (Bangla: 0.900, Marathi: 0.882, Hindi: 0.879 semantic similarity) but do not reveal its effect on downstream *inference*. To isolate dataset effects, *Regional-TinyStories* trains matched SLMs—*one per dataset variant*—holding architecture and training settings fixed while varying only the dataset (e.g., translated vs. synthetic), comparing inference-time behaviour.

Applied to Hindi (54M parameters, Sarvam-1 tokenisation; Tab. 4), all SLMs trained on our *synthetic* corpus outperforms its translated counterpart across all metrics, with substantially larger gains on content-oriented criteria—Context, Completeness, Creativity (mean $\Delta = +2.020$)—than on form-oriented Fluency and Grammar (mean $\Delta = +0.568$; Grammar is smallest; see Sec. 3.1). Because downstream utility hinges on prompt-faithful, context-aware generation, these results indicate that translation-induced semantic loss disproportionately degrades the capabilities that matter most. More broadly, *Regional-TinyStories* offers a scalable, inference-grounded protocol for *dataset evaluation*—extending beyond synthetic vs. translated to alternative recipes and translation schemes—providing actionable guidance and underscoring the value of carefully curated synthetic data in low-resource modelling.

6 Conclusion

We introduce *Regional-TinyStories*, a fast, low-cost SLM framework that isolates how language, tokenization and dataset choices affect downstream inference, enabling efficient, task-relevant ranking. To lower barriers for regional NLP, we release code, 10M+ stories, and 35+ SLM checkpoints, allowing early-stage and compute-constrained groups to run controlled ablations. While scaling narrows some gaps, we identify that achieving contextually grounded Marathi inference remains comparatively challenging under current resources. Across Hindi, Marathi, and Bangla, our inference-grounded analyses complement corpus-bound metrics (Rényi entropy, MorphScore) by directly assessing prompt-faithful generation. We find that (i) Indic-specialised tokenization yields larger gains on content-oriented metrics than the Tiktoken baseline; (ii) Sarvam and SUTRA deliver comparable downstream performance despite SUTRA's broader multilingual focus; and (iii) synthetic data consistently outperforms translated data, especially for context and completeness. Future work will extend coverage to additional languages and scripts and test whether these trends persist at larger model scales.

Additional: SLM versus GPT Inference

We compare SLMs with GPT-4 reference stories across Hindi, Marathi, and Bangla under matched prompts, revealing consistent gains through mid-scale models, cross-language gaps (Bangla strongest, Hindi steady, Marathi more volatile), and diminishing returns without better tokenization and data curation. Crucially, semantics-oriented evaluation distinguishes models more reliably than surface overlap in morphologically rich Indic settings. For full language-wise results, scale trends, and correlation analyses, see Appendix J.

Limitations

- **Unknown bias in SOTA LLMs & Tokenizers:** Employing LLM-as-a-Judge evaluations introduces ambiguity in cross-language comparisons due to unknown differences in the proficiency of state-of-the-art LLMs across languages. This uncertainty is further compounded by analogous variation in tokenizer performance

- **Poor Inter-rater-agreement for Marathi:** Our human evaluations highlight low agreement amongst rating scores for our Marathi SLMs, reducing the validity of inter-language comparisons for Marathi. This is likely due to heterogeneous fluency among evaluators. Enhancing Marathi evaluation quality is a priority for future work

- **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.

- **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:** BELU, METEOR, BERTScore and LaBSE capture semantic and lexical 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 computational 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 are committed to addressing these limitations in future work.

Impact Statement and Ethical Considerations

Impact. This work lowers the cost of systematic, inference-grounded experimentation in Indic NLP by providing a compact framework, open-source code, 10M+ stories, and 35+ SLM checkpoints. These resources enable students, early-stage researchers, and compute-constrained groups—particularly in rural or underserved regions—to run controlled ablations and engage with regional language technologies. Beyond resources, *Regional-TinyStories* serves as a practical *benchmarking tool*: it complements corpus-bound metrics (e.g., tokenization efficiency, semantic similarity) by directly measuring how design choices (tokenizers, dataset recipes, scaling) affect prompt-faithful generation. While we do not claim educational or literacy outcomes on their own, the approach can support the creation of age-appropriate materials in Hindi, Marathi, and Bangla, and can assist educators and community organizations experimenting with local-language content creation. The framework is also applicable as a preliminary screening step for future benchmarks and larger models, guiding where limited effort and funding can have the greatest downstream impact.

Ethical Considerations. The generation of children's content in regional languages requires careful stewardship. We recommend:

- *Community review and cultural stewardship:* involve native speakers, educators, and cultural practitioners in data curation and model evaluation to avoid stereotyping and to respect local norms.
- *Age-appropriate safety:* apply content filters, toxicity checks, and style constraints for child-facing outputs; clearly label AI-generated materials.
- *Bias and quality auditing:* regularly audit models for harmful biases, hallucination, and uneven performance across dialects; report known failure modes and uncertainty.

Our focus on small models reduces the computational footprint relative to large-scale alternatives, but practitioners should still disclose training settings and consider environmental costs. Overall, *Regional-TinyStories* is intended to complement—not supplant—human expertise, offering an inference-based foundation for responsible, inclusive progress in regional NLP.

References

Andrej Karpathy. 2022. nanogpt. `https://github.com/karpathy/nanoGPT`.

Catherine Arnett. 2023. Best practices for open multilingual llm evaluation. `https://huggingface.co/blog/catherinearnett/multilingual-best-practices`.

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

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.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, and Shing-Chi Cheung. 2020. Semmt: A semantic-based testing approach for machine translation systems. *CoRR*, abs/2012.01815.

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, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 437–441, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

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.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. 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.

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.

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, 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. 2022. Training compute-optimal large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '22.

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.

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

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*.

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.

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.

OpenAI. 2024. Tiktoken. `https://github.com/openai/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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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.

Melanie Revilla and Carlos Ochoa. 2017. Ideal and maximum length for respondent-friendly on-line surveys. *International Journal of Market Research*, 59(5):557–565.

Sarvam. 2024. Sarvam 1 : The first indian language llm. `https://www.sarvam.ai/blogs/sarvam-1`.

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, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

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.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Shiyue Zheng, Tuhin Ghosal, Ethan Chi, and et al. 2023b. 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 LLMA-as-a-Judge Numerical Tables

A.1 Hindi, Marathi & Bangla Hyperparameter Comparisons

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi SLMs									
64	**12**	5.00	1.057	6.374	7.227	7.390	7.959	8.450	7.480
512	**6**	53.00	0.518	7.734	7.783	7.806	8.554	8.912	8.158
1024	**7**	157.00	0.513	7.695	7.806	7.830	8.580	8.910	8.164
64 *95% CI*	-	-	-	[6.32 – 6.42]	[7.20 – 7.26]	[7.36 – 7.42]	[7.93 – 7.98]	[8.43 – 8.47]	[7.46 – 7.51]
512 *95% CI*	-	-	-	[7.70 – 7.77]	[7.75 – 7.81]	[7.79 – 7.83]	[8.53 – 8.57]	[8.90 – 8.92]	[8.14 – 8.17]
1024 *95% CI*	-	-	-	[7.66 – 7.73]	[7.78 – 7.84]	[7.81 – 7.85]	[8.56 – 8.60]	[8.90 – 8.92]	[8.15 – 8.18]
Pairwise Mann-Whitney U Test with Bonferroni Correction (α = 0.05/no. of comparisons)–*All pairs show difference at p < 0.001*									
GPT-4o-mini	-	-	-	8.915	9.725	8.493	9.803	9.835	9.354
GPT-4o	-	-	-	9.368	9.899	8.749	9.880	9.898	9.559
Gemini-1.5-flash	-	-	-	9.154	9.856	8.770	9.744	9.752	9.455
LLaMA-3.3-70B	-	-	-	9.254	9.956	8.870	9.844	9.852	9.555
Marathi SLMs									
64	**12**	5.00	2.224	6.050	7.520	7.465	8.094	8.813	7.588
512	**6**	54.00	1.662	7.154	8.127	7.902	8.854	9.146	8.236
1024	**7**	157.00	1.640	7.641	7.697	7.200	8.257	8.940	7.947
64 *95% CI*	-	-	-	[5.99 – 6.10]	[7.48 – 7.56]	[7.44 – 7.47]	[8.05 – 8.11]	[8.80 – 8.84]	[7.57 – 7.63]
512 *95% CI*	-	-	-	[7.09 – 7.22]	[8.09 – 8.17]	[7.87 – 7.93]	[8.82 – 8.88]	[9.11 – 9.16]	[8.21 – 8.28]
1024 *95% CI*	-	-	-	[7.59 – 7.69]	[7.66 – 7.74]	[7.17 – 7.23]	[8.23 – 8.28]	[8.93 – 8.95]	[7.93 – 7.96]
Pairwise Mann-Whitney U Test with Bonferroni Correction (α = 0.05/no. of comp.)–*All pairs show difference at p < 0.001; higher p-value for 54M vs 157M*									
GPT-4o-mini	-	-	-	8.630	9.459	8.416	9.359	9.441	9.061
GPT-4o	-	-	-	9.329	9.812	8.770	9.728	9.749	9.477
Gemini-1.5-flash	-	-	-	9.032	9.806	8.615	9.677	9.676	9.361
LLaMA-3.3-70B	-	-	-	9.132	9.906	8.715	9.777	9.776	9.461
Bangla SLMs									
64	**12**	5.00	1.136	6.760	7.289	7.563	7.968	8.507	7.617
512	**6**	54.00	0.569	7.507	7.645	7.693	8.420	8.816	8.016
1024	**7**	157.00	0.557	7.567	7.639	7.740	8.409	8.832	8.037
64 *95% CI*	-	-	-	[6.72 – 6.81]	[7.26 – 7.32]	[7.54 – 7.59]	[7.94 – 7.99]	[8.48 – 8.53]	[7.60 – 7.64]
512 *95% CI*	-	-	-	[7.47 – 7.55]	[7.61 – 7.68]	[7.69 – 7.71]	[8.40 – 8.44]	[8.80 – 8.83]	[8.00 – 8.04]
1024 *95% CI*	-	-	-	[7.53 – 7.61]	[7.61 – 7.67]	[7.73 – 7.76]	[8.39 – 8.43]	[8.82 – 8.85]	[7.98 – 8.05]
Pairwise Mann-Whitney U Test with Bonferroni Correction (α = 0.05/no. of comparisons)–*All pairs show difference at p < 0.001; except 54M vs 157M*									
GPT-4o-mini	-	-	-	8.953	9.692	8.637	9.695	9.734	9.342
GPT-4o	-	-	-	9.340	9.786	8.800	9.821	9.827	9.515
Gemini-1.5-flash	-	-	-	9.233	9.870	8.894	9.860	9.854	9.542
LLaMA-3.3-70B	-	-	-	9.333	9.970	8.994	9.960	9.954	9.642
Pairwise Mann-Whitney U Test with Bonferroni Correction (α = 0.05/no. of comparisons)–*across 54 pairs, 3 sizes x 6 metrics x 3 langauges* **All 54 pairs statistically different at p < 0.001 level of significance** except 3–*see Sec. H.5.4 for details*									

Table 5: **Inference scores** across key-sizes of models trained on Hindi, Marathi and Bangla synthetic TinyStories datasets — Lighter the better — Each score represents the average rating given by GPT-4o across 3,000 stories generated using different SLM configurations, as well as state-of-the-art LLMs listed in the first column.

A.2 Evaluations scores when LLaMA-3.3-70B is used as the LLM-as-a-Judge

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi SLMs									
Evlaution scores** when SLM Inference stories are evaluated **by GPT-4o									
64	**12**	5.00	1.057	6.374	7.227	7.390	7.959	8.450	7.480
512	**6**	53.00	0.518	7.734	7.783	7.806	8.554	8.912	8.158
1024	**7**	157.00	0.513	7.695	7.806	7.830	8.580	8.910	8.164
Evaluation scores** when SLM Inference stories are evaluated **by LLaMA-3.3-70B									
64	**12**	5.00	1.057	4.288	7.314	6.154	7.668	8.525	6.790
512	**6**	53.00	0.518	6.084	7.825	6.643	7.982	8.698	7.446
1024	**7**	157.00	0.513	6.217	8.052	6.876	8.208	8.921	7.655
Marathi SLMs									
*Evlaution scores when SLM Inference stories are evaluated **by GPT-4o***									
64	**12**	5.00	2.224	6.050	7.520	7.465	8.094	8.813	7.588
512	**6**	54.00	1.662	7.154	8.127	7.902	8.854	9.146	8.236
1024	**7**	157.00	1.640	7.641	7.697	7.200	8.257	8.940	7.947
Evaluation scores** when SLM Inference stories are evaluated **by LLaMA-3.3-70B									
64	**12**	5.00	2.224	3.177	6.189	5.678	7.125	8.146	6.063
512	**6**	54.00	1.662	5.131	7.454	6.395	7.757	8.780	7.104
1024	**7**	157.00	1.640	5.482	7.644	6.524	7.823	8.844	7.263
Bangla SLMs									
Evlaution scores** when SLM Inference stories are evaluated **by GPT-4o									
64	**12**	5.00	1.136	6.760	7.289	7.563	7.968	8.507	7.617
512	**6**	54.00	0.569	7.507	7.645	7.693	8.420	8.816	8.016
1024	**7**	157.00	0.557	7.567	7.639	7.740	8.409	8.832	8.037
Evaluation scores** when SLM Inference stories are evaluated **by LLaMA-3.3-70B									
64	**12**	5.00	1.136	5.091	7.359	6.574	7.735	8.567	7.065
512	**6**	54.00	0.569	6.187	7.679	6.763	7.962	8.645	7.447
1024	**7**	157.00	0.557	6.385	7.836	6.977	8.111	8.841	7.630

Table 6: **Inference evaluation scores** when SLM generated stories are evaluated by GPT-4o versus LLaMA-3.3-70B. *While a difference in the absolute values of scores between the two LLM-as-Judges is observed, **key trends are maintained***

B Inference Evaluation and Factual Comparison

B.1 SLM Inference and Evaluation Methodology

This section explores the end-to-end prompting-to-evaluation procedure for our SLMs.

B.1.1 Preparing Prompts for SLM Inference

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/` in our repository)[11]. 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 generated three distinct stories, resulting in a total of 3000 distinct stories per SLM configuration, which are evaluated using the LLM-as-a-Judge paradigm detailed below.

B.1.2 SLM Inference & Evaluation

Our SLMs are causal models, generating one token at a time, following a given input phrase/prompt.

We evaluated the story generation quality across Hindi, Marathi and Bangla at varying model parameter sizes.

Each model was provided with 1,000 prompts, with the corresponding 3,000 generated stories scored across five evaluation categories: context awareness, completeness, grammar, fluency, and creativity using a discrete rating scale (1–10). Generated stories are evaluated by GPT-4o using the prompt below. The overall score is reported as the average of the 5 evaluation metrics.

Evaluation Pormpt 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*}".

Only return a JSON dictionary in the following format:

```
{
"context awareness": "...",
"completeness": "...",
"grammar": "...",
"fluency": "...",
"creativity": "...",
}
```

B.2 Cross-Lingual Inference Comparison

We observe a consistent pattern of context drift in the Marathi outputs. As shown in Tab. 7, 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 Tabs. 9, 10, and 11.

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 Tab. 8, 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. H). 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.

[11] Link To: Code & Datasets

.

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 7: **Inference comparisons** across 54M parameter models trained on our Hindi, Marathi and Bangla TinyStories datasets.

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 बिस्तर 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 सांगितले, "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, "पिंकी, 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 सांगितले, "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 8: 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 जताई 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 9: **Inference comparisons** of SLMs trained on our **Hindi** TinyStories dataset 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 चिरंतन 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 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*
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 10: **Inference comparisons** of SLMs trained on our **Marathi** TinyStories dataset 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 খিরও 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 ফিরল। 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 11: **Inference comparisons** of SLMs trained on our **Bangla** TinyStories dataset across key model sizes.

C Human Evaluation of SLM Generated Stories

C.1 Introduction

While automatic metrics provide scalable evaluation methods for language models, human evaluation remains the gold standard for assessing generation quality, particularly for morphologically complex and resource-constrained languages. This study presents a systematic human evaluation framework applied to small language models trained on Hindi, Bengali, and Marathi, contributing empirical evidence for model scaling effects in multilingual contexts. We employ both raw score analysis for direct performance comparison against GPT-4 baselines and normalised score analysis to account for systematic evaluator biases, providing a comprehensive understanding of model capabilities.

C.2 Methodology

C.2.1 Experimental Design

Overview We conduct a large-scale human evaluation of short-story inferences produced by three model sizes (5M, 53M, 157M parameters) across Hindi, Marathi, and Bangla. The design closely follows the recommendations of van der Lee et al. (2019)–a leading ACL published resource on best-practice guidelines for human evaluation of Automatically Generated Text, complemented by Arnett (2023), recent recommendations on human-centred study procedures for multilingual LLMs Elangovan et al. (2024), and survey-length constraints shown to reduce respondent fatigue (Revilla and Ochoa, 2017). Our goals are (i) to obtain reliable, reader-oriented quality estimates that complement automatic metrics and (ii) to verify whether GPT-4o's automatic judgments align with human preferences.

Evaluation Procedure Because no existing benchmark targets child-friendly prompt continuations in the three target languages, we adopt a *task-matched* survey. Each rater sees the *exact* evaluation prompt previously given to GPT-4o and scores every story on the same metrics–*Grammar*, *Fluency*, *Creativity*, *Completeness*, and *Context* using a 10-point Likert scale (1 = *Very Poor*, 10 = *Perfect*). Clear definitions of these metrics and one calibrated example are provided to minimise bias and confusion.

Participants We target a broad pool of adult native speakers recruited via social media communities and our personal networks. To achieve adequate statistical power, each language collects ratings from at least $N \geq 40$ unique raters. Raters self-report language proficiency so that analyses can be weighted or filtered by fluency.

Materials and Workload Control 3 unique Google Form is created per language, each following the same design/template. A Google Form used is available at the end of this Appendix in Sec C.4. Each form contains *one* story from each model size (total 3 stories–5M, 54M, 157M) to keep completion time below the 5–10 minute window that preserves data quality in online surveys (Revilla and Ochoa, 2017). A total of three unique forms per language were created to ensure story diversity (9 unique stories evaluated by humans per language).

Quality Assurance To detect inattentive responses, every form includes a single attention-check item ("If you are reading this, select "). Submissions failing the check or containing implausibly fast completion times are excluded.

Statistical Analysis For each dimension, we compute mean ±95 % confidence interval (CI) and deviation of human evaluations from GPT-4o ratings. Statistical significance was assessed using the Kruskal-Wallis test. Raw scores underwent z-score normalisation within evaluators to control for individual rating bias. Inter-rater reliability is quantified via Krippendorff's α; values $\alpha > 0.67$ are interpreted as substantial agreement.

Ethical Considerations Participation is voluntary; no personal data beyond self-reported fluency and email IDs (to maintain a count on unique participants, as two people can have identical names) is stored. All procedures adhere to the institutional review guidelines for minimal-risk online surveys.

C.3 Results

C.3.1 Participants

We conduct a relatively large-scale Human Evaluation cohort, involving 100+ unique evaluators and 650+ story instances evaluated across languages.

Given our task of evaluating simple short stories meant for 5-to-7-year-old children, we consider evaluators across three categories–Native (Native/-Bilingual Fluency), Advanced (Full Professional

Language	Story Evaluations	#Evaluators	Fluency Excl. Rate	Native	Advanced	Working
Hindi	282	45	–	24	13	28
Bengali	246	42	1.2%	66	16	–
Marathi	219	47	2.7%	34	19	20

Table 12: Breakdown of **evaluation counts and evaluator proficiency** by language.
Participants with elementary proficiency or below were excluded.

Language	1st Place	2nd Place	3rd Place	Kruskal-Wallis H	p-value
Bengali	54M (0.783)	157M (0.679)	5M (0.514)	12.802	0.002
Hindi	157M (1.389)	54M (1.188)	5M (0.716)	29.638	< 0.001
Marathi	157M (0.538)	54M (0.071)	5M (-0.609)	252.532	< 0.001

Table 13: Model **Performance Rankings** by Language (Normalised Scores).
Values in parentheses represent mean z-score normalised performance (μ).
*All comparisons show **statistically significant differences** between model sizes.*

Language	Model	Raw Mean (SD)	Min-Max	Normalized μ
Hindi	157M	7.50 (1.85)	1-10	1.389
	54M	7.54 (1.85)	1-10	1.188
	5M	6.77 (2.34)	1-10	0.716
Bengali	54M	6.74 (2.05)	1-10	0.783
	157M	6.58 (2.09)	1-10	0.679
	5M	6.32 (2.19)	1-10	0.514
Marathi	157M	6.53 (2.19)	1-10	0.538
	54M	5.61 (2.45)	1-10	0.071
	5M	4.37 (2.14)	1-10	-0.609

Table 14: **Raw Score Performance** Summary.
Raw scores represent average ratings across all five evaluation dimensions.
Normalised μ values show z-score adjusted performance rankings within each language

Language	Overall α	Interpretation	Context Awareness	Completeness	Creativity	Grammar	Fluency
Bengali	0.705	Strong	0.748	0.662	0.733	0.719	0.663
Hindi	0.666	Good	0.632	0.554	0.610	0.723	**0.810**
Marathi	0.332	Poor	0.367	0.226	0.477	0.338	0.252

Table 15: **Inter-rater Reliability** Analysis–*Krippendorff's alpha*–
Reliability thresholds: ≥ 0.8 *(Excellent),* ≥ 0.67 *(Good),* ≥ 0.6 *(Acceptable),* < 0.6 *(Poor).*

Proficiency) and Working (Limited Working Proficiency). We report total number of evaluations belonging to each category (last 3 columns) along with a breakdown of evaluators in Tab. 12.

C.3.2 Analysis of Z-score Normalised Evaluations

Raw scores underwent *z-score* normalisation within evaluators to control for individual rating bias. For each evaluator i, individual scores were transformed using:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{1}$$

where x_{ij} represents the raw score for evaluation j by evaluator i, μ_i is evaluator i's mean score across all evaluations, and σ_i is evaluator i's standard deviation.

C.3.3 Statistical Significance Analysis

Statistical significance was assessed using the *Kruskal-Wallis* test, chosen due to:

- Non-normal distributions (Shapiro-Wilk tests, $p < 0.05$)
- Heteroscedasticity across model groups
- Ordinal nature of Likert scale ratings

Results detailed in Tab. 13.

Model	Context	Completeness	Creativity	Fluency	Grammar	Overall
5M						
Mean Score	6.32	6.25	6.06	7.49	7.72	6.77
95% CI	[5.72, 6.93]	[5.62, 6.88]	[5.50, 6.63]	[6.99, 7.99]	[7.33, 8.12]	[6.29, 7.25]
Δ vs. GPT-4o	-2.68	-2.75	-0.94	-0.51	-1.28	-1.63
54M						
Mean Score	7.20	7.51	7.23	7.86	7.88	7.54
95% CI	[6.72, 7.68]	[7.01, 8.01]	[6.79, 7.67]	[7.47, 8.26]	[7.47, 8.28]	[7.15, 7.92]
Δ vs. GPT-4o	-1.80	-1.49	+0.23	-0.14	-1.12	-0.86
157M						
Mean Score	7.31	7.03	6.93	8.14	8.10	7.50
95% CI	[6.66, 7.96]	[6.28, 7.79]	[6.21, 7.65]	[7.49, 8.79]	[7.63, 8.57]	[6.98, 8.03]
Δ vs. GPT-4o	-1.69	-1.97	-0.07	+0.14	-0.90	-0.90

Table 16: **Human Evaluation** Performance of SLMs trained on **Hindi** TinyStories dataset – *Higher Scores are better – 9 unique stories human evaluated per model size*

Model	Context	Completeness	Creativity	Fluency	Grammar	Overall
5M						
Mean Score	4.25	4.33	4.59	4.33	4.36	4.37
95% CI	[3.73, 4.76]	[3.84, 4.82]	[4.11, 5.07]	[3.85, 4.80]	[3.86, 4.86]	[3.94, 4.80]
Δ vs. GPT-4o	-4.75	-4.67	-2.41	-3.67	-4.64	-4.03
54M						
Mean Score	5.36	5.59	5.63	5.82	5.66	5.61
95% CI	[4.78, 5.94]	[5.02, 6.16]	[5.09, 6.17]	[5.22, 6.42]	[5.13, 6.18]	[5.13, 6.09]
Δ vs. GPT-4o	-3.64	-3.41	-1.37	-2.18	-3.34	-2.79
157M						
Mean Score	6.48	7.08	6.26	6.67	6.15	6.53
95% CI	[5.99, 6.96]	[6.59, 7.58]	[5.73, 6.79]	[6.18, 7.16]	[5.66, 6.65]	[6.09, 6.96]
Δ vs. GPT-4o	-2.52	-1.92	-0.74	-1.33	-2.85	-1.87

Table 17: **Human Evaluation** Performance of SLMs trained on **Marathi** TinyStories dataset – *Higher Scores are better – 9 unique stories human evaluated per model size*

Model	Context	Completeness	Creativity	Fluency	Grammar	Overall
5M						
Mean Score	5.99	5.99	6.12	6.74	6.74	6.32
Δ vs. GPT-4o	-3.01	-3.01	-2.54	-1.26	-0.92	-2.15
95% CI	[5.48, 6.50]	[5.47, 6.51]	[5.66, 6.59]	[6.32, 7.17]	[6.32, 7.17]	[5.90, 6.73]
54M						
Mean Score	6.55	6.76	6.35	7.13	6.93	6.74
Δ vs. GPT-4o	-2.45	-2.24	-2.31	-0.87	-0.74	-1.72
95% CI	[6.10, 7.00]	[6.31, 7.20]	[5.87, 6.84]	[6.73, 7.53]	[6.50, 7.36]	[6.35, 7.14]
157M						
Mean Score	6.26	6.45	6.35	6.93	6.93	6.58
Δ vs. GPT-4o	-2.74	-2.55	-2.31	-1.07	-0.74	-1.88
95% CI	[5.74, 6.78]	[5.99, 6.92]	[5.90, 6.81]	[6.53, 7.33]	[6.52, 7.33]	[6.20, 6.97]

Table 18: **Human Evaluation** Performance of SLMs trained on **Bangla** TinyStories dataset – *Higher Scores are better – 9 unique stories human evaluated per model size*

C.4 Showcased Example of Our Google Form

Email *

☐ Record as the email to be included with my response

What would you rate your fluency in Hindi? *

○ No proficiency

○ Elementary proficiency (basic greetings, simple phrases)

○ Limited working proficiency (basic work/social conversations)

○ Full professional proficiency (advanced, detailed discussions with rare minor errors)

○ Native or bilingual fluency

Evaluation Guideline / Rubric

The given Hindi short stories are **meant for 5-7-year-old children**. Keeping in mind the target demographic, rate the stories 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 specified prompt.
- Evaluate *completeness* as how well the story covers all parts — beginning, middle, and end — to form a full, clear narrative.
- Evaluate *creativity* as how original, interesting, and imaginative the story is in terms of ideas, themes, and writing style.
- Evaluate *grammar* as how correct and appropriate the language is, including sentence structure and use of grammar.
- Evaluate *fluency* as how smoothly and naturally the story reads, with clear and easy-to-follow sentences.

Figure 4: **Introduction & Evaluation Guidelines** of one of our Hindi Google Forms

Example Story and Evaluation

Prompt:
एक दोपहर, एक जिज्ञासु गिलहरी ने एक गुप्त पेड़घर का रास्ता दिखाया।

Story:
एक दोपहर, एक जिज्ञासु गिलहरी ने एक गुप्त पेड़घर का रास्ता दिखाया। उस पेड़ के पास एक बड़ा सा बक्सा था, जिसमें ढेर सारे रंग-बिरंगे खिलौने थे। गिलहरी ने सोचा कि उसे उस बक्से को खोलना चाहिए।

गिलहरी ने अपनी छोटी सी पूंछ हिलाते हुए अपने दोस्तों को बुलाया। उसने कहा, "चलो, हम सब मिलकर इस बक्से को खोलते हैं!" सभी दोस्तों ने उत्सुकता से सहमति जताई। गिलहरी ने एक छोटी सी लकड़ी की टहनी ली और उसे बक्से के ठंडे पत्तों पर चिपकाया। सभी दोस्त एक साथ मिलकर टहनी पर खड़े हो गए।

जब उन्होंने टहनी को खींचा, तो बक्सा धीरे-धीरे खुलने लगा। अचानक, बक्से से एक सुंदर गुड़िया बाहर गिर पड़ी। गुड़िया ने कहा, "धन्यवाद, दोस्तों! मैं इस गुड़िया के लिए बहुत खुश हूँ।" सभी ने गुड़िया को देखा और फिर हंसने लगे। गिलहरी ने गुड़िया को अपने दोस्तों से मिलवाया और सभी ने मिलकर खेलना शुरू कर दिया।

उस दिन, गिलहरी और उसके दोस्त बहुत मज़ा कर रहे थे। उन्होंने समझा कि साथ रहकर हर काम आसान हो जाता है। जब सूरज ढलने लगा, तो सभी ने एक-दूसरे को अलविदा कहा और अपने-अपने घर लौट गए। गिलहरी ने सोचा कि हर दिन कुछ नया सीखने का मज़ा ही अलग है।

Evaluation:
context awareness: 8
completeness: 9
grammar: 9
fluency: 9
creativity: 8

Important
Please provide **fair and honest evaluations**. The purpose of this study is not to obtain "perfect" scores, but to capture your genuine impressions of the stories.

Next Page 1 of 4 Clear form

Figure 5: **Example Story Evaluation** for reviewer calibration in one of our Hindi Google Forms

Hindi Story (Code: 5423)

Prompt: ऊँचे बादलों में, एक दोस्ताना ड्रैगन ने उड़ान भरी।

Story:
ऊँचे बादलों में, एक दोस्ताना ड्रैगन ने उड़ान भरी। राधा और मोहन ने डरने की बजाय, हिम्मत दिखाई। उन्होंने मिलकर ड्रैगन को भगाने का फैसला किया। उन्होंने एक योजना बनाई और शोर मचाते हुए, जंगल में चले गए।

आखिरकार, राधा और मोहन ने ड्रैगन को डराकर भगा दिया। गाँव में सब लोग तालियाँ बजाने लगे। राधा और मोहन ने बादलों के नीचे एक नई कहानी बनाई। उस दिन की बातें और खेलों ने सभी को एक नई ऊर्जा दी।

Evaluation Guideline / Rubric
(same as before)

The given Hindi short story is **meant 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 specified prompt.
- Evaluate *completeness* as how well the story covers all parts — beginning, middle, and end — to form a full, clear narrative.
- Evaluate *creativity* as how original, interesting, and imaginative the story is in terms of ideas, themes, and writing style.
- Evaluate *grammar* as how correct and appropriate the language is, including sentence structure and use of grammar.
- Evaluate *fluency* as how smoothly and naturally the story reads, with clear and easy-to-follow sentences.

Figure 6: **Sample story to evaluate (Part-I)** in one of our Hindi Google Forms

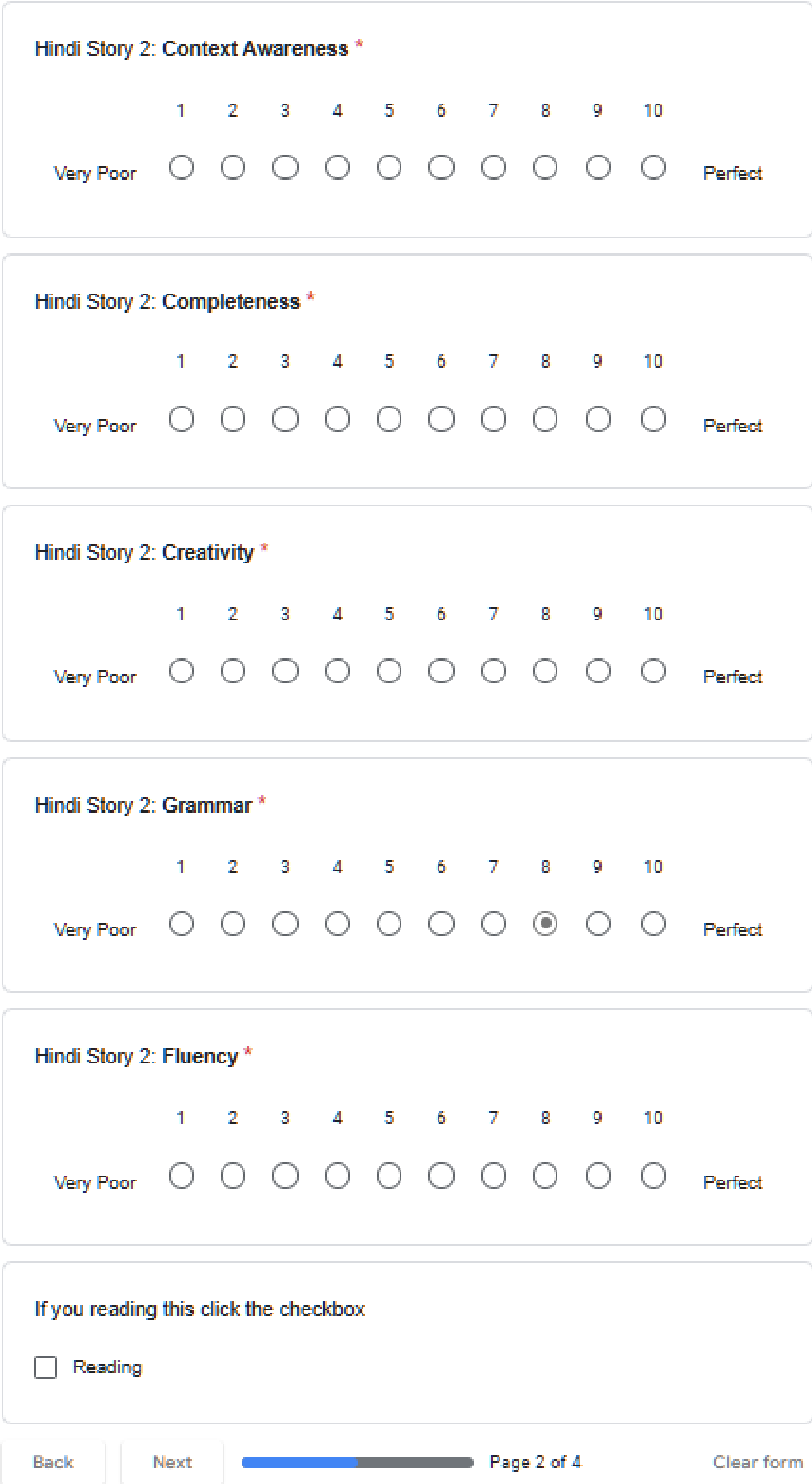


Figure 7: **Sample story to evaluate (Part-II)** in one of our Hindi Google Forms

D SLM Architecture Training & Tokenizer Details

The implementation for this section is available at `training-inference/` in our repository[12].

D.1 nanoGPT based SLM Architecture

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

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

$$y = x + \mathrm{MHA}(x), \tag{2}$$

$$z = y + \mathrm{MLP}(\mathrm{LN}(y)), \tag{3}$$

$$\mathrm{Block}(x) = \mathrm{LN}(z). \tag{4}$$

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** (Details in next Sec. D.4). 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 ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\lambda{=}0.1$).
  - **Total Steps**: 5000
  - **Minimum LR**: Cosine decay, after Warmup Steps (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 Steps**: 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 Random Seed (`torch.manual_seed`) used for training and inference: `1337`. Exact reproducibility (seeds) and training configurations can be found and easily customised through (`training-inference/train.py`, `training-inference/sample.py`) found in our repository.

D.2 Training Details

Training Regime

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

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 19: **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.*

[12]Link To: Code & Datasets

D.3 Training (WANDB) Logs

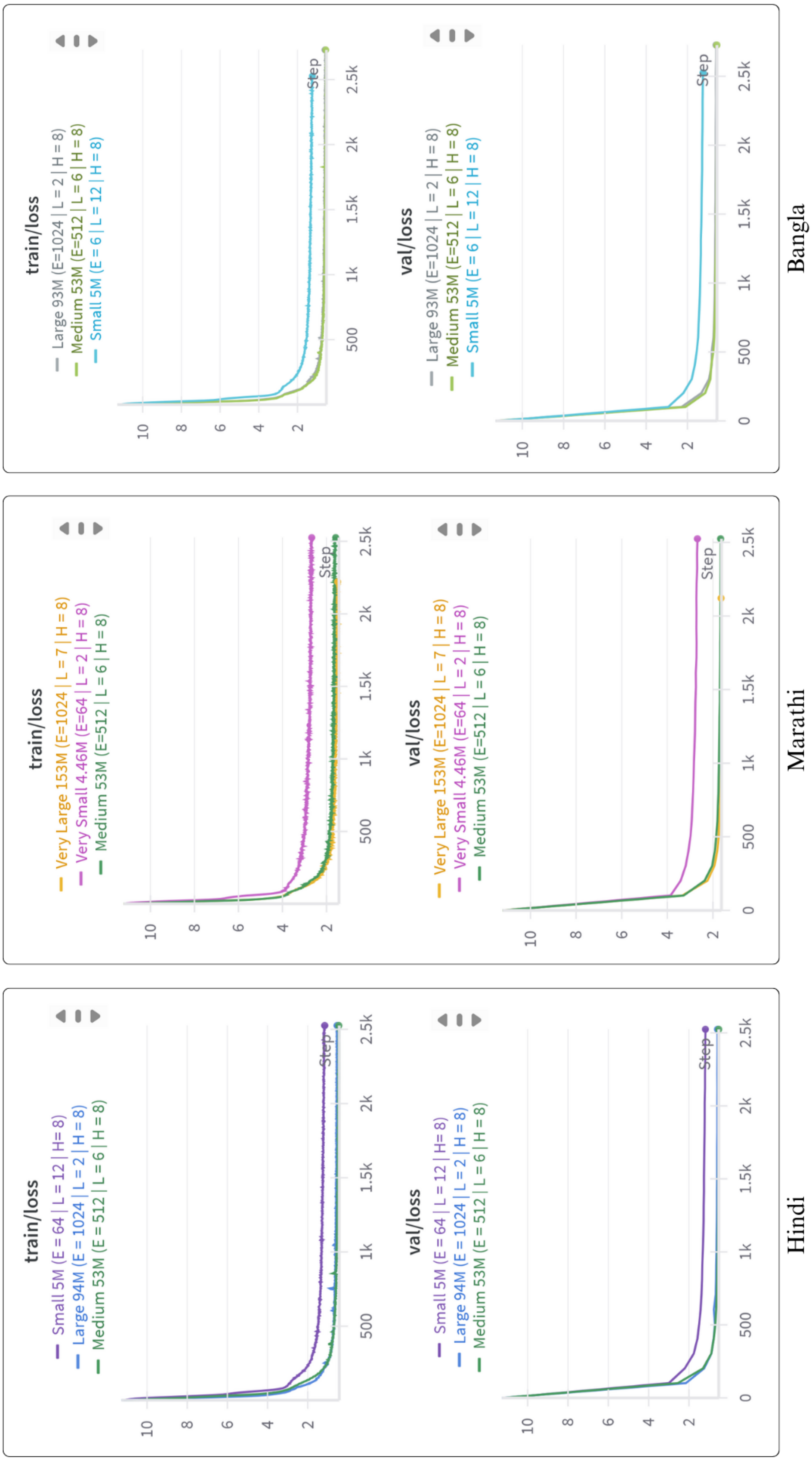


Figure 8: Training & Testing loss curves for Small, Medium and Large Models across Hindi, Marathi & Bangla

D.4 Tokenizer Details

D.4.1 Overview of Tokenizers

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 20: Total Token Counts across Datasets for Each Language by Tokenizer

Vocabulary Size

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

Underlying Framework

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

D.4.2 Tokenizer Overview: Sarvam, SUTRA and Tiktoken

All three tokenizers 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 ~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 ~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× more tokens per word in Indic scripts, inflating latency and memory footprints (Petrov et al., 2023).

E Synthetic Dataset Generation through LLM Prompting

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

E.1 Nouns, Verbs, Adjectives & Features in Prompts

TinyStories (Eldan and Li, 2023b) presents a method for constructing a dataset of English short stories by issuing a large number of prompts to GPT-4. Each prompt is generated by sampling a noun, verb, adjective, and "feature," which are then inserted into a predefined prompt template designed to elicit age-appropriate short story generation from the model.

In line with the seminal work, our prompt generation process utilised language-specific lexical resources, each containing 300 nouns, 300 verbs, and 300 adjectives curated for children aged 5-7. We extended the original work by developing custom 'feature' lists with explicitly positive and age-appropriate narrative elements, such as themes of friendship and kindness, while excluding inappropriate content like violence found in the seminal work. These features (e.g., "playfully mysterious," "colorfully gentle") guide the emotional tone and moral direction of the generated stories.

Our curated nouns, verbs, adjectives and feature lists are publicly available (See folder `prompting/prompt_gen/<language>` in our repository)[13].

E.2 Prompt Generation Algorithm

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

This approach effectively prevents repetition patterns (avoidance of 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.

[13]Link To: Code & Datasets

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}
  $ID \leftarrow$ ConcatenateIndices(n, v, a, f)
  $TripletID \leftarrow$ ConcatenateIndices(n, v, a)
  if $ID \notin UsedIDs$ and $TripletID \notin UsedTriplets$ **then**
    $UsedIDs \leftarrow UsedIDs \cup \{ID\}$
    $UsedTriplets \leftarrow UsedTriplets \cup \{TripletID\}$
    $prompt \leftarrow$ FormatTemplate(n, v, a, f)
    $Prompts \leftarrow Prompts \cup \{prompt\}$
  else
    $DuplicateCount \leftarrow DuplicateCount + 1$
  end if
end while
return $Prompts, DuplicateCount$

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

Utilising each prompt template, we ask GPT-4o-mini to generate 1000 stories which are then evaluated by GPT-4o; the results (Fig. 9) are used to determine the optimal prompt as described in the next subsection.

E.4 Optimal Prompt Template

Based on the aforementioned cross-prompt evaluation (Fig. 9), the Level 2+ template, which produced the best results across languages, followed

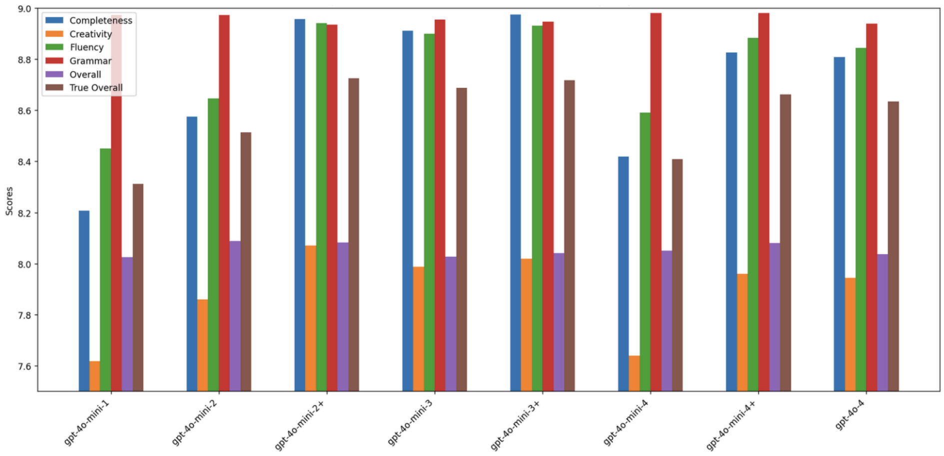


Figure 9: **Prompt Complexity Comparison**
gpt-4o-mini-1: Model used = GPT-4o-mini, Prompt complexity = 1
Prompt complexity = 2+ (where + indicates increased word limit of 350–500 words)

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

the structure below

This template's effectiveness stems from several critical elements:

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

E.5 Implementation and Data Generation

For each language, we generated 2 million unique prompts (see `prompting/prompt-gen/create_prompts.py` in our repository). These prompts were subsequently supplied to GPT-4o-mini, which was selected for its optimal balance of cost-effectiveness, multilingual capabilities, and robust API support.

To manage this large-scale task, the story generation process was highly parallelized. By deploying a configuration of 4 concurrent API sessions with 16 threads each on a 16 vCPU system, we achieved a generation throughput of approximately 100 stories per minute (see `prompting/request_helper.py` in our repository).

The final output of this scalable pipeline is a dataset comprising 2 million synthetic stories for each of the target languages: Hindi, Bangla, and Marathi. This end-to-end framework addresses the limitations of the foundational work by ensuring both linguistic diversity and quality. It thereby provides a crucial resource for training effective Small Language Models in these under-represented regional languages.

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

Implementation for this section can be found at `analysis/` in our repository[14].

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.

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 21: 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 22: English TinyStories ROUGE scores sample

However, when the same methodology was applied to the entire Bangla TinyStories dataset, ROUGE uniformly produced zero values:

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

Table 23: Average ROUGE F1 Scores (Bangla)

F.1.2 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.

This finding highlights an extreme limitation of ROUGE—particularly for Indic languages—where the lack of exact n-gram overlap, as reflected in zero ROUGE scores, hampers meaningful analysis of corpus diversity. This underscores the need for regionally relevant evaluation metrics better suited to linguistic variation in these languages.

[14]Link To: Code & Datasets

F.2 Language-Aware Analysis of Training Data

In the absence of ROUGE-based analysis to measure lexical diversity in our datasets, we employ popular language-aware metrics suitable for measuring both semantic and lexical diversity in the case of non-English text.

F.2.1 Overview of the Metrics

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) measures n-gram lexical similarity between candidate and reference translations. Scores range from 0 (no overlap) to 1 (perfect match). BLEU favors shorter texts and may underestimate semantic equivalence.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005), improving on BELU, combines precision and recall, incorporates stemming and synonym matching, and penalizes disordered word sequences. While also measuring lexical similarity, it often correlates better with human judgments than BLEU, especially for languages with flexible word order.

BERTScore (Zhang et al., 2020) computes similarity using contextual embeddings from pretrained language models. By comparing token embeddings rather than exact matches, BERTScore captures paraphrases and synonyms, making it more robust for semantically evaluating generated text.

F.2.2 Lexical and Semantic Analysis of Training Data

For each language, we randomly sample two stories—one designated as reference—and compute BLEU and METEOR for lexical overlap, and BERTScore for semantic similarity.

BLEU Score Bangla story pairs yield low lexical overlap (mean BLEU = 0.078, σ = 0.126; range 0.003–0.421) (Fig. 11). Only one pair (index 5) shows moderate alignment (BLEU = 0.421). The generally low BLEU scores corroborated our ROUGE findings, confirming significant lexical diversity in the dataset.

METEOR Score METEOR sits between BLEU and BERTScore (mean = 0.153, σ = 0.046; range 0.071–0.231) (Fig. 12), reflecting its balance of precision, recall, and linguistic matching. These low to intermediate METEOR scores further highlight the lexical diversity in our dataset.

BERTScore Semantic similarity remains very high (mean = 0.967, σ = 0.012; range 0.944–0.982) (Fig. 13). This indicates that despite diverse wording, generated Bangla stories preserve semantic meaning and narrative themes, likely because each story is fabricated by GPT-4o-mini based on themes and narratives suited for 5-7 year-old children. Qualitative examples in the next section (Sec. F.2.4) support this finding.

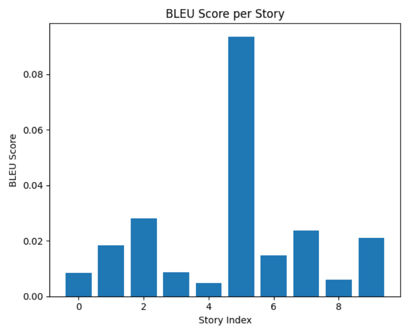


Figure 11: **BLEU** for 10 random Bangla stories.

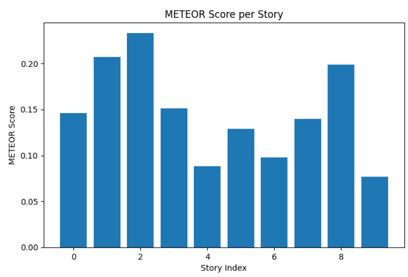


Figure 12: **METOR** for 10 random Bangla stories.

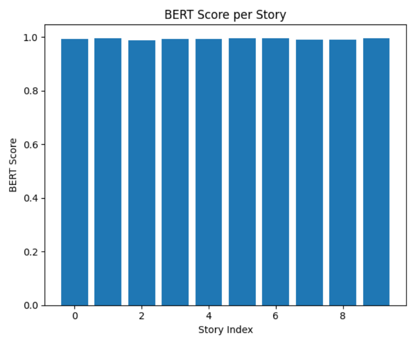


Figure 13: **BERT** for 10 random Bangla stories.

F.2.3 Co-relation Analysis

The moderate correlation between BLEU and METEOR ($r = 0.63$) suggests that despite ME-

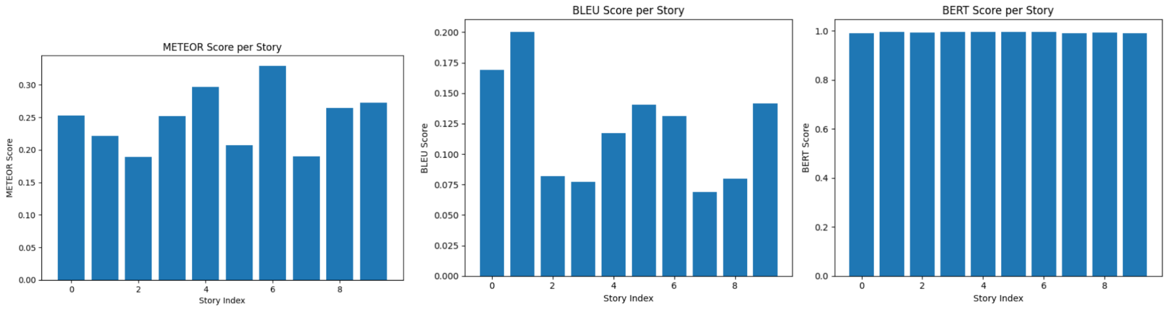


Figure 14: BLEU, BERT & METOR Scores for 10 random **Hindi** stories.

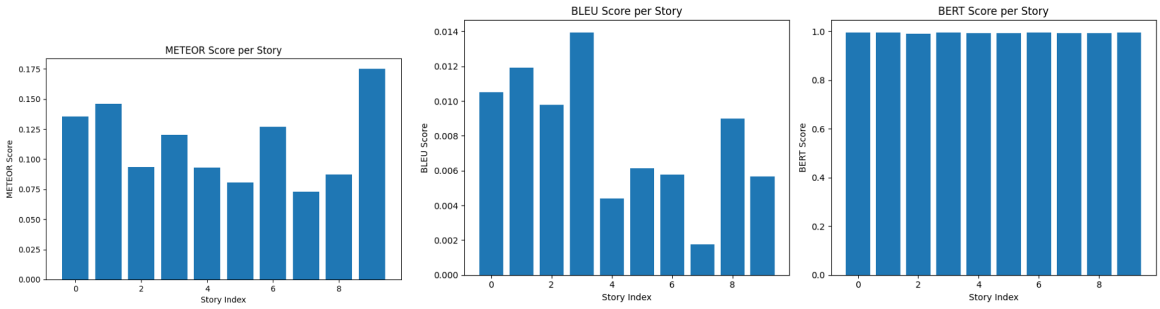


Figure 15: BLEU, BERT & METOR Scores for 10 random **Marathi** stories.

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

Table 24: Pearson Correlation Coefficients Between Metrics for Bangla Stories

TEOR'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 similar observations hold true for randomly sampled synthetic training data in Hindi (Fig. 14) and Marathi (Fig. 15).

F.2.4 Case Study: Qualitative Analysis of Bangla Story Pairs

To illustrate the difference between lexical overlap and semantic similarity, we present a Bangla story pair from our dataset, along with the English translations (Fig. 16 & Fig. 17).

Although both stories depict a child's outdoor experience, they diverge lexically—in vocabulary, characters, and setting—yet BERTScore still reports high semantic similarity (0.961), reflecting their shared narrative elements and emotional tone.

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. While we conduct preliminary analyses, based on our research, we propose future research along the following directions:

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:

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 16: 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 17: A story for reference from the same dataset

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

F.4 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 datasets reveals lexical diversity while semantic similarity:

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

Our findings confirm *our generation approach produces semantically coherent content with lexical diversity, rather than relying on exact phrase repetition*, while demonstrating how traditional lexical metrics fail to capture semantic equivalence in morphologically rich Indian languages.

On a conclusive note, the pitfalls of traditional metrics and underrepresented research towards finding Regional language-aware metrics, we motivate researchers to explore new, rigorous and relevant semantic and lexical metrics or an evaluation pipeline along the ideas we outline in Sec. F.3.

G Analysis of Translated Datasets

G.1 LaBSE-based Translation Quality Assessment

To thoroughly understand the loss in semantic information via translation, we utilise Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022).

LaBSE represents an optimal choice for evaluating translation quality in our multilingual context for several compelling reasons. Unlike traditional lexical metrics such as BLEU and METEOR, which rely heavily on n-gram overlap and surface-level matching, LaBSE captures deep semantic similarity through its multilingual sentence embeddings trained on 109 languages. This semantic understanding is particularly crucial for morphologically rich Indian languages, where the same meaning can be expressed through vastly different surface forms.

While BERT-based metrics like BERTScore offer semantic evaluation capabilities, they typically require language-specific models or multilingual variants that may not be equally optimised across all target languages. LaBSE's architecture, specifically designed for cross-lingual tasks, ensures balanced representation quality across Hindi, Bengali, and Marathi without the bias toward high-resource languages that affects many multilingual models. Furthermore, LaBSE's training on parallel data makes it inherently suited for translation evaluation, as it learns to map semantically equivalent sentences across languages into similar embedding spaces.

G.1.1 Evaluation Methodology

We translated the TinyStories corpus into Hindi, Bengali, and Marathi and measured translation fidelity by computing cosine similarity between each English story and its translation. From 10,000 stories per language, we randomly sampled 1,000 story pairs. To assess statistical confidence, we applied bootstrap resampling with 10,000 iterations, reporting 95% confidence intervals for the mean similarity scores.

G.1.2 Translation Quality Results

Our LaBSE analysis reveals significant variations in translation quality across the three target languages:

Bengali achieved the highest mean similarity score of 0.9002 (95% CI: [0.8988, 0.9016]), demonstrating superior semantic preservation during translation. This aligns with Bengali's relatively straightforward morphological structure compared to other Indo-Aryan languages, facilitating more accurate machine translation.

Marathi exhibited intermediate performance with a mean similarity of 0.8824 (95% CI: [0.8808, 0.8838]),

Hindi showed the lowest scores at 0.8793 (95% CI: [0.8778, 0.8807]). The narrow confidence intervals (CI widths: Bengali 0.0028, Hindi 0.0029, Marathi 0.0030) indicate high precision in our estimates.

Statistical significance testing based on non-overlapping confidence intervals confirms that Bengali translations significantly outperform both Hindi and Marathi ($p < 0.05$), while Marathi significantly exceeds Hindi. These differences, though statistically significant, represent meaningful variations in translation fidelity that directly impact downstream model performance.

G.1.3 Distribution Analysis and Quality Indicators

Shapiro-Wilk tests revealed non-normal distributions for all languages ($p < 0.001$), suggesting heterogeneous translation quality within each corpus. The interquartile ranges (IQRs)—Hindi: 0.0312, Bengali: 0.0303, Marathi: 0.0311—indicate consistent spread in translation quality, while the minimum scores (Hindi: 0.7966, Bengali: 0.7691, Marathi: 0.7488) reveal occasional translation failures that could introduce significant noise during model training.

Notably, Bengali achieved perfect similarity (1.0) for some translations, while Hindi and Marathi peaked at 0.9895 and 0.9977, respectively. This ceiling effect in Bengali suggests that certain simple narrative structures translate nearly perfectly, while more complex constructions show greater variation.

G.1.4 Implications for Model Performance

These LaBSE findings provide crucial context for understanding the performance gap between models trained on translated versus synthetic data. The correlation between translation quality scores and model performance metrics is striking:

1. **Hindi models** trained on translated data achieved only 6.30 overall inference score compared to 8.16 with synthetic data—a

22.8% performance degradation that aligns with Hindi's lowest LaBSE scores.

2. **Bengali models**, despite having the highest translation quality, still showed an 11.6% performance drop (7.08 vs 8.02) when using translated data, suggesting that even high-quality translations introduce artefacts that impede language modelling.

3. **Marathi models** exhibited a 14.8% decline (7.06 vs 8.30), consistent with their intermediate translation quality.

The universal performance degradation, even for high-quality Bengali translations, supports our hypothesis that translation-induced noise fundamentally differs from the natural linguistic patterns in synthetic data. LaBSE scores, while useful for assessing semantic fidelity, cannot capture subtle grammatical inconsistencies, unnatural collocations, or cultural misalignments that emerge during translation—factors that significantly impact next-token prediction accuracy.

G.1.5 Translation Quality as a Dataset Selection Criterion

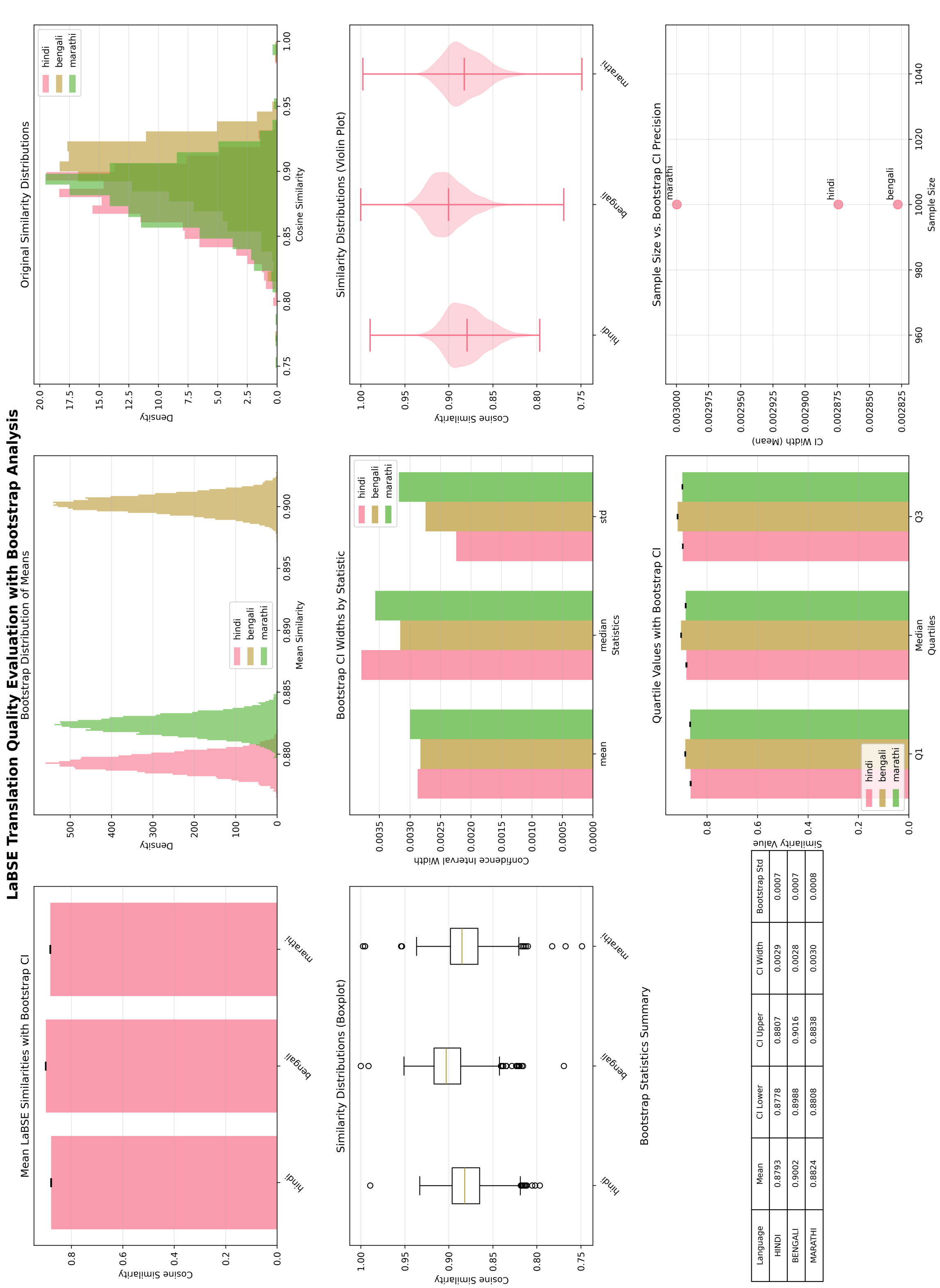
Our analysis establishes LaBSE evaluation as a valuable analysis step for translated datasets. The strong negative correlation between translation quality variance (measured by IQR and standard deviation) and model performance suggests that consistency in translation quality may be as important as mean quality scores. For resource-constrained scenarios where synthetic data generation is infeasible, we recommend:

1. Filtering translated stories based on LaBSE thresholds (e.g., > 0.85 similarity)

2. Prioritising languages with narrower confidence intervals for more predictable outcomes

3. Augmenting high-quality translations with small amounts of synthetic data, following Boughorbel et al. (2024)

This multi-faceted evaluation approach—combining LaBSE semantic similarity with downstream inference metrics by utilising our Regional-TinyStories framework for dataset comparisons—provides a comprehensive framework for assessing and selecting training data for low-resource language modelling.



Language	Mean	CI Lower	CI Upper	CI Width	Bootstrap Std
HINDI	0.8793	0.8778	0.8807	0.0029	0.0007
BENGALI	0.9002	0.8988	0.9016	0.0028	0.0007
MARATHI	0.8824	0.8808	0.8838	0.0030	0.0008

Figure 18: Bootstrap Analysis with LaBSE to Quantify Semantic Loss in Machine Translation

H Statistical Analysis of Inference Results and SLM *Langauge-learning* analysis

The implementation for this section is available at `results/compare_models_statistics.py` in our repository[15].

H.1 Limitations

Although identical procedures are used to prepare data and train the Hindi, Marathi, and Bengali models, cross-lingual comparison of inference quality—evaluated using GPT-4o—remains inherently limited. This is due to the unknown relative proficiency of GPT-4o across different languages, as well as potential inconsistencies in the Sarvam tokenizer's handling of distinct scripts. These factors introduce external biases and noise that may affect the reliability of cross-language performance comparisons.

To mitigate these limitations, we undertake detailed statistical analyses of the GPT-4o evaluation scores (presented in this section) and further validate the use of LLM-as-a-Judge through human evaluation of inference outputs across languages (see Sec. C).

H.2 Overview of Analyses

Motivation To uncover trends and patterns in the evaluation scores (see Fig 3 & Tab. 5) generated using GPT-4o as an LLM-as-a-Judge for the 3,000 stories produced (see Sec. B) by each of our 54M-parameter (E:512, L:6) SLMs trained on our Hindi, Marathi, and Bengali synthetic TinyStories datasets, we conduct comprehensive statistical analyses.

Hierarchical Emergence of Capabilities across Performance Metrics Examining the hierarchical emergence of capabilities via mean, median, and standard deviation metrics; assessing parameter elasticity; analysing correlations between evaluation dimensions; and performing performance gap analysis across languages and models. These analyses offer deeper insight into the behavioural and performance characteristics of the models across multiple evaluation criteria.

Statistical Significance Analysis Utilising distribution-relevant metrics, we measure and report the statistical significance of our inter-language tests.

[15] Link To: Code & Datasets

Next sections—Hierarchical Emergence of Capabilities

H.3 Optimal Model Configuration

The inflexion 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. Interestingly, the 157M (E: 1024, L: 7) parameter models perform similarly across all metrics to the optimal 54M model for all 3 languages, despite its ~3x larger parameter count.

Note. Unless specified otherwise, our primary model of analysis remains the 54M (E: 512, L: 6) configuration trained on our synthetic TinyStories datasets for each language.

H.4 Hierarchical Emergence of Capabilities

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

H.4.1 Overview of Trends

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: 8.85, Bangla: 8.42), completeness (Hindi: 7.78, Marathi: 8.12, Bangla: 7.64), and trailed by context awareness (Hindi: 7.73, Marathi: 7.15, Bangla: 7.51).

This pattern aligns with the developmental progression observed in the seminal TinyStories research, highlighting that grammatical competence emerges earlier than contextual understanding, regardless of language.

H.4.2 Inverse Relationship & Metric-Specific Patterns

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 suggests that as performance on a particular aspect improves, consistency also increases.*

Model	Metric	Mean	95% CI	Median	Std Dev
Hindi	Context Awareness	7.73	[7.70–7.77]	8.00	1.01
	Completeness	7.78	[7.75–7.81]	8.00	0.86
	Creativity	7.81	[7.79–7.83]	8.00	0.58
	Fluency	8.55	[8.53–8.57]	9.00	0.56
	Grammar	8.91	[8.90–8.92]	9.00	0.34
Marathi	Context Awareness	7.15	[7.09–7.22]	8.00	1.42
	Completeness	8.12	[8.09–8.17]	7.00	1.06
	Creativity	7.90	[7.87–7.93]	8.00	0.69
	Fluency	8.85	[8.82–8.88]	8.00	0.64
	Grammar	9.14	[9.42–9.47]	9.00	0.50
Bangla	Context Awareness	7.51	[7.47–7.55]	8.00	1.11
	Completeness	7.64	[7.61–7.68]	7.00	0.85
	Creativity	7.69	[7.67–7.71]	8.00	0.59
	Fluency	8.42	[8.40–8.44]	8.00	0.59
	Grammar	8.82	[8.80–8.83]	9.00	0.42

Table 25: **Statistical Properties of Evaluation Metrics**——*for 54M (E:512, L:6) models trained on Hindi, Marathi, and Bangla TinyStories datasets*

Metric	Hindi	Marathi	Bangla	Average
Context Awareness	1.01	1.42	1.11	1.18
Completeness	0.86	1.06	0.85	0.92
Creativity	0.58	0.69	0.59	0.62
Fluency	0.56	0.64	0.59	0.60
Grammar	0.34	0.50	0.42	0.42
Overall	0.52	0.67	0.57	0.59
Average	0.65	0.87	0.69	0.70

Table 26: **Standard Deviation Comparison** Across Metrics and Languages—*for 54M (E:512, L:6) models trained on Hindi, Marathi and Bangla synthetic TinyStories datasets*

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 *it may represent a more advanced capability that remains challenging even as models improve.*

H.4.3 Consistent Hierarchy of Variability

Across all languages, Grammar exhibits both the highest mean scores and the lowest standard deviations (average SD: 0.42), indicating that grammatical competence is a foundational capability that, once learned, is consistently retained across models. In contrast, Context Awareness shows the lowest mean scores and the highest standard deviation (average SD: 1.18), highlighting substantial variability in the model's ability to maintain contextual coherence.

These patterns suggest that while grammatical accuracy stabilises early in model development, context-sensitive reasoning remains a more complex and unevenly acquired capability.

H.4.4 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.

H.4.5 Form vs. Content Metrics

Metrics related to linguistic form (grammar, fluency) consistently show lower standard deviations

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 27: **Performance Gaps** Between Metrics (Difference in Mean Scores)—*for 54M (E:512, L:6) models trained on Hindi, Marathi and Bangla synthetic TinyStories datasets*

(0.42, 0.60) than those related to content (context, completeness, creativity) (1.18, 0.92, 0.62) (see Tab. 26).

These results further suggest that form-related capabilities may develop more uniformly and easily compared to content-related capabilities.

H.4.6 Performance Gap Analysis

Metric Specific Gaps The substantial gaps between form-related metrics (grammar, fluency) and content-related metrics (context, completeness, creativity) (See Tab. 27) *highlight the models' stronger capabilities in producing structurally correct text versus semantically coherent narratives.*

On the other hand, 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*.

Consistent Gap Hierarchy Across all languages, the largest performance gap is between grammar and context awareness (average: 1.32) (See Tab. 27), 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.*

H.4.7 Conclusion—Hierarchical Emergence of Capabilities

Form-oriented capabilities (grammar and fluency) emerge as foundational proficiencies–developing in tandem, achieving high mean scores with low variability and modest parameter elasticity, while content-oriented skills (context awareness, completeness, creativity), especially context awareness, exhibit lower means, greater dispersion, and stronger sensitivity to model scale.

This consistent hierarchy across languages underscores the relative ease of mastering linguistic form versus the inherent complexity of contextual reasoning.

Next sections—Statistical Significance Analysis

H.5 Statistical Significance Analysis for Inter-language Comparisons

To assess the statistical significance of all observed cross-linguistic trends, differences and analyses in the previous sections, we conduct a thorough distribution-relevant analysis.

H.5.1 Methods

Overview We evaluated the story generation quality of three multilingual language models (Hindi, Bengali, Marathi) across three parameter scales: 5M, 54M, and 157M. Analysing pairwise, this results in a total of 6x3=18 comparisons.

Data Distribution Analysis The score distributions exhibited severe departures from normality across all model sizes. Score concentrations were extreme: at 5M (68.3% at scores 7.0–8.0), at 54M (72.4% at scores 8.0–9.0), and at 157M (74.8% at scores 8.0–9.0), creating heavily right-skewed distributions with pronounced ceiling effects. This discrete, ordinal rating scale (1–10) with extreme skewness fundamentally violates the continuous, normally distributed data assumption required for ANOVA.

H.5.2 ANOVA Violations

The data exhibited multiple critical violations across all parameter sizes: severe non-normality (evident from the extreme skewness), heteroscedasticity (unequal variances between groups), and the large sample sizes (53,988–54,000 observations) that would make any trivial deviation highly "significant," necessitating focus on effect sizes rather than p-values alone.

H.5.3 Methodological Rationale

We adopted non-parametric methods specifically designed for ordinal data: *Kruskal-Wallis* tests for comparing multiple groups without distributional assumptions, *Mann-Whitney U-tests* for robust pairwise comparisons, *Bonferroni correction* to control familywise error rate across 18 pairwise comparisons per model size ($\alpha = 0.0167$), rank-biserial correlation for effect sizes appropriate for ordinal data ($|r| \geq 0.10$, 0.30, and 0.50 representing small, medium, and large effects), and median-based reporting as appropriate measures for skewed, ordinal data. This approach avoids the inflated Type-I error rates and spurious "significance" that would result from applying parametric methods to fundamentally non-parametric data.

H.5.4 Results

Cross-Size Statistical Significance The claim that *"all pairwise differences reach significance ($p < 0.001$)" almost always holds true*:

- **5M Models**: **17/18 comparisons significant (94.4%)** — Hindi vs Bengali fluency was not significant ($p = 0.608$, $r = -0.007$)
- **54M Models**: **18/18 comparisons significant (100.0%)** ✓
- **157M Models**: **16/18 comparisons significant (88.9%)** — Hindi vs Marathi in completeness ($p = 0.820$) and overall ($p = 0.692$) were not significant

Main Effects Kruskal-Wallis tests revealed significant main effects for both model and category factors across all parameter sizes (all $p < 0.001$), with varying effect magnitudes: Model effects H-statistics ranged from 275.2 (157M) to 1649.0 (54M), while category effects consistently showed large H-statistics (16,067–19,333), *confirming that both language and evaluation dimension significantly influence performance rankings.*

H.5.5 Conclusion—Statistical Significance Analysis

The extreme skewness and ceiling effects in our ordinal data invalidate ANOVA, necessitating Kruskal–Wallis and Mann–Whitney U tests with Bonferroni correction. These tests confirmed significant main effects of model size and evaluation dimension (all $p < 0.001$).

Pairwise comparisons were significant in 94.4%, 100%, and 88.9% of cases for the 5M, 54M, and 157M models, respectively. Effect sizes ranged from small to large, underscoring meaningful language- and scale-dependent differences.

This validates our non-parametric framework for rigorously analysing the statistical significance of non-normal, ordinal scores.

I Rényi Entropy–Based Tokenizer Analysis

I.1 Implementation of Byte Premium Scaling and Statistical Analysis of Rényi entropy

In our tokenizer performance analysis, we controlled for dataset-size disparities across languages by implementing byte-premium scaling as proposed by (Arnett and Bergen, 2025). Arnett and Bergen argue that observed cross-language performance gaps often reflect inequities in training data volume, rather than inherent morphological complexity. The "byte premium" accounts for differing encoding efficiencies: identical semantic content can require varying numbers of bytes depending on orthography.

To ensure a fair comparison, we normalised each language corpus by its byte size rather than by document or word count, thereby equating information content across languages. Our processing pipeline comprises the following steps:

- Compute the total byte size of each raw corpus.
- Derive sampling weights to match a common byte budget.
- Extract tokeniser-specific subcorpora using these weights.
- Train and evaluate SLMs on the normalised datasets.

I.2 Quantifying Tokenization Efficiency via Rényi Entropy

To quantitatively assess tokenization efficiency, we computed the normalised Rényi entropy ($\alpha = 2.5$) for each language–tokenizer pair, as recommended by (Arnett and Bergen, 2025). Rényi entropy captures the uniformity of the token frequency distribution, making it a robust metric for tokenization quality.

For each pair, we measured entropy across five data splits and then calculated the mean Rényi entropy, its variance, and the corresponding 95% confidence interval. The aggregated results are presented in Table 28.

I.3 Statistical Significance via Welch's t-Test

To determine if the observed differences in mean Rényi entropy across the languages were statistically significant, we performed pairwise Welch's t-tests, which are appropriate as they do not assume equal variances between the samples (see Table 29).

Extremely small p-values ($p \ll 0.01$) for all pairwise comparisons confirm that the Rényi entropy differences among Hindi, Marathi, and Bengali are statistically significant for both the `sutra` and `sarvam` tokenizers. The large magnitudes of the corresponding t-statistics underscore the substantive size of these effects. Even after normalizing corpus size via byte-premium scaling, these results demonstrate that tokenizer efficiency varies meaningfully across the three languages. This rigorous analysis underpins our paper's conclusions regarding comparative tokenizer performance.

I.4 Entropy Overview

Vocabulary Context. Table 30 reports Rényi entropy for Hindi, Marathi, and Bangla under two Indic tokenisers. SARVAM-1 ($\sim$ 68K vocab) is markedly leaner than SUTRA ($\sim$ 256K), a difference reflected in their entropy profiles.

Marathi Dominance at Low α. At $\alpha = 0.5$ (rare-token focus), Marathi tokenised with SARVAM-1 reaches 10.926—the highest value across all settings—quantitatively confirming the language's long-tail morphology (Dolamic and Savoy, 2010; Arnett and Bergen, 2025).

I.4.1 Rare–Common Token Spectrum

$\alpha = 0.5$. For all three languages, SARVAM-1 yields higher entropy than SUTRA, indicating superior sensitivity to rare subwords.

$\alpha \geq 1.0$. The trend reverses: SUTRA consistently exceeds SARVAM-1. For Bangla at $\alpha = 1.0$, entropy climbs from 8.649 (SARVAM-1) to 8.745 (SUTRA); at $\alpha = 2.5$, Marathi shows a gap of 1.16 points (7.670 vs. 6.506).

Trade-off. SARVAM-1 excels at modelling rare variants, whereas SUTRA's larger vocabulary captures richer mid- and high-frequency structure.

I.4.2 Morphological Interpretation

Agglutinative Marathi naturally exhibits higher entropy than fusional Bangla and mixed Hindi. The elevated low-α values underscore Marathi's suffix-heavy word formation and validate the need for tokenisers with morphology-aware vocabularies.

I.5 Conclusion

Byte-premium scaling to neutralise dataset-size disparities, enabling a rigorously controlled comparison of tokenizer efficiency across Hindi, Marathi,

Language	Tokenizer	Mean Entropy	Variance ($\times 10^{-7}$)	95% CI Lower	95% CI Upper
Hindi	SUTRA	7.100 685	0.915	7.100 309	7.101 060
Hindi	Sarvam	6.279 743	1.400	6.279 278	6.280 208
Marathi	SUTRA	7.677 890	2.860	7.677 225	7.678 554
Marathi	Sarvam	6.505 551	0.539	6.505 263	6.505 839
Bengali	SUTRA	7.319 165	1.110	7.318 751	7.319 578
Bengali	Sarvam	6.303 569	0.733	6.303 233	6.303 905

Table 28: **Mean Rényi entropy, variance, and 95% CI** by language and tokenizer.

Tokenizer	Pair	t-statistic	p-value
SUTRA	Hindi–Marathi	-2099.895446	<0.005
SUTRA	Hindi–Bengali	-1085.998694	<0.005
SUTRA	Marathi–Bengali	1272.845495	<0.005
Sarvam	Hindi–Marathi	-1145.624005	<0.005
Sarvam	Hindi–Bengali	-115.269686	<0.005
Sarvam	Marathi–Bengali	1266.364085	<0.005

Table 29: **Pairwise Welch's t-test** results for Rényi entropy differences.

Language	Tokenizer	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$	$\alpha = 2.5$
Hindi	Sarvam	10.071	8.520	7.462	6.755	6.285
	SUTRA	9.744	8.624	7.919	7.439	7.101
Marathi	Sarvam	10.926	9.370	8.151	7.189	6.506
	SUTRA	10.416	9.290	8.571	8.059	7.670
Bangla	Sarvam	9.923	8.649	7.660	6.870	6.300
	SUTRA	9.790	8.745	8.080	7.630	7.319

Table 30: **Rényi entropy** (H_α) for Hindi, Marathi, and Bangla SLMs across Sarvam and SUTRA tokenizers—*calculated using Zouharvi's Tokenization Scorer (GitHub Repository)*

and Bengali. Normalising at the byte level ensured that any observed differences stemmed from tokeniser design and language morphology rather than corpus volume.

Using the normalised Rényi entropy ($\alpha = 2.5$) as an information-theoretic proxy for tokenisation quality, we revealed consistent, statistically significant gaps among the three languages ($p \ll 0.01$ via pairwise Welch's t-tests). Even after scaling, SUTRA and SARVAM-1 diverged markedly: SARVAM-1 achieved higher entropy at low α, signalling superior coverage of rare subwords, while SUTRA excelled at $\alpha \geq 1.0$, capturing mid- and high-frequency structure through its larger vocabulary.

Language morphology modulated these effects. Agglutinative Marathi exhibited the highest entropy—especially under SARVAM-1 at $\alpha = 0.5$—validating long-tail suffixation patterns, whereas fusional Bangla and mixed Hindi showed lower values. These findings corroborate prior work (Arnett and Bergen, 2025) and underscore that tokenizer architecture must be attuned to a language's morphological profile.

In sum, (i) byte-level corpus equalisation isolates tokenizer behaviour, (ii) Rényi entropy provides a sensitive yardstick for efficiency across the rare–common token spectrum, and (iii) significant, morphology-driven differences persist despite corpus normalisation. Future research should explore adaptive or hybrid tokenisation strategies that dynamically balance rare-token coverage with mid-frequency robustness, particularly for morphologically rich languages.

J Additional Benchmarking of SLM vs GPT-4 stories

J.1 Lexical vs Semantic Metrics

We evaluated SLM story generation using lexical (BLEU, METEOR) and semantic metrics (LaBSE, COMET), revealing useful discriminative differences. For e.g. in Hindi SLMs, lexical metrics showed moderate variance—BLEU 13.4%, METEOR 10.5%—across models spanning 4.46M-153M parameters, while semantic metrics demonstrated contrasting patterns: COMET showed higher variance (15.5%) but LaBSE showed minimal variance (3.7%). This suggests that while SLMs achieve similar topical alignment (LaBSE), they differ in overall semantic quality (COMET), which is detailed with story examples in Sec. B.

The strong LaBSE-COMET correlation (0.723) versus the negative BLEU correlations (-0.268 with LaBSE and -0.453 with COMET) in Hindi validate that n-gram metrics fail to capture the quality differences that semantic metrics tend to detect in morphologically rich Indian languages. The negative correlations suggest that lexical metrics may even be inversely related to semantic quality.

Performance improves with parameter count up to 40-60M parameters, then plateaus. For Hindi models, semantic quality (COMET-DA) shows substantial gains from 0.56 to 0.66 (15.5% variance), while topical alignment (LaBSE) quickly saturates at 0.74-0.77 with minimal variance (3.7%). This pattern, where SLMs rapidly acquire topical relevance but continue to improve in narrative coherence with scale, is confirmed by both LLM judges and human evaluations. The plateau beyond 50M parameters suggests that tokenizer quality and data curation, rather than raw scale, become the primary drivers of story quality in Indic language models, establishing semantic evaluation as essential for these languages.

J.2 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 contextualised representations and compute semantic similarity relative to reference texts or prompts. However, COMET-DA remains primarily tuned for translation adequacy and fluency, and underrepresents qualities central to narrative generation, such as discourse structure, stylistic nuance, and imaginative coherence. Additionally, encoder variance and normalisation 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 significantly better with human preferences than lexical n-gram metrics, such as 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.

J.3 Conclusion

The following observations are made based on the results in Tables. 31 to 36:

- 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 exhibit an upward trend in COMET-DA scores, albeit 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 modelling 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 31: 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 32: 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 33: 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 34: 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 35: 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 36: LaBSE similarity for Bangla checkpoints (ascending size).