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

Anonymous ACL submission

Abstract

We extend the English-only TinyStories paradigm to three Indic languages—Hindi, Marathi, and Bangla—by releasing 2 M synthetic and translated stories per language and training 5–157 M-parameter Small Language Models (SLMs). We introduce the *Regional-TinyStories framework*, which treats SLMs as cost-effective stand-ins for Large Language Models (LLMs), enabling rapid, variable-wise inference analysis. Using this framework we (i) chart *language-learning* dynamics, finding performance to plateau at 54 M parameters and to follow a stable Grammar → Fluency → Completeness → Context hierarchy; (ii) rank tokenizers, showing the Indic–specific Sarvam-1 to outperform SUTRA and the generic Tiktoken across all metrics; and (iii) demonstrate that synthetic corpora yield up to 30% higher context scores than machine-translated data. Results are validated with an LLM-as-Judge ensemble (GPT-4o and LLaMA-3.3-70B) and a 100 +-participant human study, which replicates key trends while revealing systematic score inflation in automated evaluations. Our findings establish Regional-TinyStories as a lightweight, end-to-end pipeline for benchmarking tokenizers, datasets, and SLM designs in low-resource settings, offering a reproducible path toward scalable, context-faithful generation in regional languages.[1].

1 Introduction

Research on Language Models (LMs) has predominantly focused on scaling to multi-billion parameters (Brown et al., 2020; Chowdhery et al., 2022), as larger models tend to offer improved performance despite increased computational demands (Hoffmann et al., 2022b). Recently, Small Language Models (SLMs) have surged to the fore with the BabyLM challenge (Warstadt et al., 2023), which frames language learning under tight data budgets. Follow-up work shows their promise: Muckatira et al. (2024) demonstrate zero-shot gains from tiny vocabularies, while Hugging Face anchors this shift—TinyLlama (1.1 B params, 1 T tokens) matches much larger baselines through rigorous curation (Zhang et al., 2024). Indic efforts push further: Nemotron-Mini-Hindi-4B (Joshi et al., 2025) and the 2B-param Sarvam-1 (Sarvam AI, 2024) beat 8–9 B baselines on IndicXTREME and TriviaQA, proving sub-5 B models can deliver state-of-the-art accuracy when paired with tailored tokenizers and synthetic Indic text.

The TinyStories framework by Eldan and Li (2023a) explores a unique research tangent by demonstrating that English SLMs with fewer than 50M parameters can perform meaningful short-story inference when trained on smaller, carefully curated datasets. This mirrors children's language acquisition, typically exposed to fewer than 100 million words by age 13 (Gilkerson et al., 2017). TinyStories establishes that coherent text generation can emerge from smaller architectures, language capabilities develop hierarchically, and model width correlates with knowledge retention while depth enhances contextual understanding. While today's "small" 5B parameter models remain heavy, ultra-compact and culturally aligned SLMs hold a realistic path to sovereign AI and broad digital inclusion—yet expanding TinyStories efforts to regional languages remains few ((Boughorbel et al., 2024a)). Motivated by this, we extend and improve the TinyStories paradigm to regional languages, thereby amplifying underrepresented linguistic communities in NLP, reducing the entry barrier for Small Language Models (SLMs), whilst uncovering their potential as tools for comparative analyses.

The ultimate measure of a language model's success is the quality of its inference, as this capability underpins its utility in real-world applica-

[1] Link To: Code & Datasets

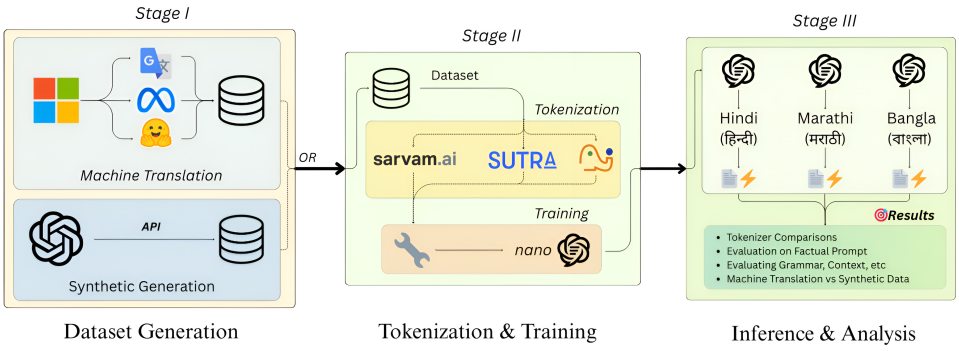


Figure 1: **Regional-TinyStories Framework**–Training distinct SLMs for variable-wise inference analysis.

tions. Therefore, the most rigorous method for quantifying the downstream effects of fundamental components—such as tokenization schemes, dataset compositions, or the morphological complexity of a language—involves training a unique Large Language Model (LLM) from scratch for each experimental variable and subsequently evaluating its inference performance. However, the prohibitive computational and financial costs associated with training contemporary LLMs render such an exhaustive, end-to-end approach fundamentally inefficient for systematic comparative analysis. To circumvent this, **we propose a framework that utilises the ultra-compact TinyStories-based SLMs as a time and cost-efficient proxy for LLMs**, facilitating the rapid training of a distinct SLM for each variation. This enables a detailed and direct comparison of how foundational choices impact inference-time metrics, thereby providing critical insights that can guide the more resource-intensive development of their larger counterparts, in line with established scaling principles (Kaplan et al., 2020; Hoffmann et al., 2022a; Hu et al., 2020; Tal et al., 2022).

Leveraging our proposed framework–*Regional-TinyStories* (Fig.1), we extend the TinyStories paradigm to Hindi, Marathi, and Bangla-three Indic languages with substantial speaker populations and distinct linguistic features. By **applying this framework to conduct a series of targeted comparative analyses**, our work makes the following **key contributions:**

- Generalising the TinyStories paradigm to any Regional language (this study focuses on Hindi, Marathi and Bangla), detailing effective data preparation and pre-training methodologies to achieve high inference quality with smaller models.
- Open-sourcing code for our Regional-TinyStories framework, 10M+ synthetic and translated stories across the three Indian languages and weights for 30+ SLMs[2].
- Uncovering trends in Inference performance dynamics for varying model sizes and validating these claims with examples, statistical analyses and Human Evaluations.
- Establishing a methodology for comparing *linguistic complexity* (defined later).
- A fine-grained comparative analysis of multiple tokenizers, analysis via traditional tokenization efficiency metrics and contrasting a legacy tokenizer (Tiktoken) with specialised Indic tokenizers such as Sarvam and SUTRA.
- Demonstrating that training on synthetic datasets vastly outperforms translated datasets for inference quality.

2 Methodology: Data Generation, Training and Evaluation Experiments

2.1 Training Data Preparation

Our research extends the TinyStories framework to Hindi, Marathi and Bangla by training SLMs on data generated by either: (i) Translating the original English TinyStories dataset (Eldan and Li, 2023b) into Indian languages, or (ii) Generating synthetic data using SOTA LLMs (See Stage-I in Fig. 1).

2.1.1 Synthetic Data

Our data generation pipeline, applied uniformly across Hindi, Marathi, and Bangla, incorporates several procedural enhancements that overcome key shortcomings of the seminal TinyStories methodology for data generation.

The process begins with the rigorous curation and manual verification of a high-quality vocabu-

[2]Link To: Code & Datasets

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 Prompt Template**.

lary, consisting of 300 nouns, verbs, and adjectives in addition to carefully curated 'features' (e.g., "playfully mysterious," "colorfully gentle") which guide the emotional tone and moral direction of the generated stories for each language (Sec. E.1).

Utilising this vocabulary in conjunction with an optimised prompt template–shown in Fig. 2 (Left) (details in Secs. E.3, E.4), we generate a corpus of 2 million prompts. The uniqueness of each prompt is guaranteed by a custom duplicate avoidance algorithm, a critical step for ensuring data diversity (Sec. E.2).

Finally, this set of unique prompts is provided as input to *GPT-4o-mini to synthesise the corresponding 2M stories that form our final synthetic dataset (Sec. E.5) for each language*, which we open-source[3].

2.1.2 Translated Data

Utilising previously established approaches for translation to Indic languages (Doshi et al., 2024), (Boughorbel et al., 2024a), we translated the complete TinyStories dataset (~2M stories) from English to each of Hindi, Marathi and Bangla using NLLB-200-3B (Team et al., 2022) and Google Translate.

2.2 Training Data Evaluation

We perform a robust language-aware analysis of our training data utilising conventional metrics, LLM-as-a-judge assessments, and human evaluations to quantify lexical diversity and semantic coherence across languages.

2.2.1 Synthetic Data

Conventional Metrics Our analysis exposes the limitations of ROUGE (Lin, 2004), manifesting as the "Zero-ROUGE Phenomenon" (zero scores across all stories; see Sec. F.1). Consequently, we sample story pairs from the training set, evaluating pair-wise lexical similarity with BLEU and METEOR, and semantic similarity with BERTScore (metric desc. in Sec. F.2.1). We observe low lexical overlap (BLEU: $\bar{x} = 0.078$, METEOR: $\bar{x} = 0.153$), indicating high lexical diversity, alongside high semantic alignment (BERTScore: $\bar{x} = 0.967$), reflecting coherent and theme appropriate narratives—likely due to GPT-4o-mini's guided theme-based generation for 5–7-year-old audiences.

Additionally, when comparing a random training story with an SLM-generated story, we observe the same trends: low BLEU and METEOR scores confirm that SLMs do not simply memorise training data, while high BERTScore demonstrates that generated stories preserve age-appropriate semantics and thematic coherence relative to the training set.

LLM-as-a-Judge We leverage GPT-4o for a large-scale evaluation of our training corpus. Using the prompt template shown in Fig. 2 (Right), GPT-4o rates 3,000 randomly sampled stories across—Context, Creativity, Completeness, Fluency, and Grammar. We observe consistently strong performance ($\bar{x} \geq 8.5/10$) across all dimensions, closely matching the GPT-4o-mini results reported for each language in Table **??**.

Human Evaluation We strengthen our evaluation by manually reviewing over 300 training stories in multiple languages, ensuring each is high-quality, thematically appropriate, and free of harmful content.

2.2.2 Translated Data

Machine translation is often accompanied by semantic loss (Cohn-Gordon and Goodman, 2019). To rigorously quantify this loss, we conduct detailed analyses using conventional yet advanced semantic similarity techniques.

[3]Link To: Code & Datasets

Semantic-aware evaluation across languages To quantify how faithfully our translated TinyStories corpora preserve semantics, we employ Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2022), whose multilingual encoder captures cross-lingual semantics beyond surface n-gram overlap. LaBSE is therefore better suited than lexical metrics (BLEU, METEOR) or general-purpose multilingual BERT variants, whose optimisation is typically skewed toward high-resource languages. Using the procedure described in Sec. G.1.1, we computed cosine similarity between 1,000 English–translation story pairs per language, drawing bootstrap-based 95% confidence intervals.

Key findings LaBSE scores reveal statistically significant differences in translation fidelity (Sec. G.1.2). Bengali translations achieve the highest mean similarity (0.9002), followed by Marathi (0.8824) and Hindi (0.8793); non-overlapping confidence intervals confirm all pairwise gaps at $p < 0.05$ (See Fig. 18). Distributional inspection (Sec. G.1.3) shows heavy tails and Shapiro–Wilk non-normality, indicating heterogeneous quality and occasional failures in translation (minima as low as 0.75). These language-specific semantic discrepancies directly impact the inference performance of SLMs trained on the corresponding translated data. We further quantify this performance degradation using our Regional-TinyStories framework in the Results Section (Sec. **??**). .

2.3 SLM: Build, Train, Inference & Evaluate

Build & Train Our SLMs are built from scratch in PyTorch using the nanoGPT codebase (Karpathy, 2022), implementing decoder-only transformers with 8 attention heads at various embedding dimensions (64, 512, 768) and depths (2, 6, 12) leading to an array of parameter sizes from 5M to 157M (Sec. D.1). All models are trained for 5000 steps with 2.5% of 2M stories (one dataset) reserved for testing (Sec. D.2).

Tokenization We employ two open-source, Indic tokenizers—Sarvam-1 (Sarvam, 2024) and SUTRA-mlt-256-v2 (Bendale et al., 2024), while as a baseline, we use OpenAI's widely adopted *Tiktoken* (OpenAI, 2024). Tokenizer details are provided in Sec. D.4. Through comprehensive analysis, we identify Sarvam-1 as the most effective tokenizer: it consistently outperforms SUTRA across all evaluation metrics (see Sec. D.4.2) while offering significantly faster tokenization and a vocabulary nearly four times smaller. *Accordingly, we adopt Sarvam-1 as the default tokenizer for all subsequent analyses unless stated otherwise.*

Inference Following training, each SLM was used to generate three unique stories for each of 1,000 manually curated prompts (unseen during training), resulting in a total of 3,000 stories per SLM configuration. For evaluation, we adopt the LLM-as-judge framework (Eldan and Li, 2023a; Zheng et al., 2023), using GPT-4o to assess each story based on the evaluation prompt shown in Fig. 2 (Right). Each story is rated on a scale of 1–10 across five dimensions: Context, Completeness, Creativity, Fluency, and Grammar, with an overall score computed as the average of these five. The final reported score for each metric is the average across all 3,000 stories generated for a given SLM configuration (details in Sec. B.1.1).

3 Results & Discussions

3.1 Extending seminal TinyStories to Regional Languages

The seminal TinyStories work is confined to English. We extend this framework to our language-agnostic framework–Regional-TinyStories by employing SLMs as efficient proxies for LLMs for inference-based comparisons.

This study covers three Indian languages—Hindi, Marathi, and Bangla—and provides open-source translated and synthetically generated datasets of 2M stories per language, along with the weights of 35+ SLMs in total.[4]

3.2 Reference Tables

Tab. 1 and its extensions (Sec. A) provide all SLM scores and Mann–Whitney U tests (Bonferroni-corrected) used in subsequent analyses.

3.3 Performance Dynamics

Optimal Configuration

Performance saturates at 54M parameters; this configuration, trained with the Sarvam-1 tokenizer, serves as our default (2.3).

Hierarchical Emergence of Capabilities

- **Metric Order:** Grammar ($\bar{x} \approx 9.0$) > Fluency (≈ 8.6) > Completeness (≈ 7.9) > Context (≈ 7.5) across all languages (H.4).

[4] Link To: Code & Datasets

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
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
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*									
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.7.4 for details*									

Table 1: **Inference scores** across key-sizes of models trained on Hindi, Marathi and Bangla TinyStories synthetic datasets — Lighter the better — 3000 stories per configuration evaluated. *For detailed tables and comparisons see Appendix Sec. A*

- **Variability:** High means correlate with low dispersion (Grammar $\sigma \approx 0.42$ vs. Context $\sigma \approx 1.18$; Tab. 28).
- **Parameter Elasticity:** Grammar improves 12% and Context 33% when scaling from 4.46M to 157M parameters (H.4.4).
- **Form vs. Content:** Form metrics show tighter bands ($\sigma \leq 0.60$) than content metrics ($\sigma \geq 0.62$) (H.4.5).
- **Performance Gaps:** Largest gap Grammar–Context ($\Delta \approx 1.32$); smallest Grammar–Fluency ($\Delta \approx 0.46$) (H.4.6).

Implication

Form-oriented skills (grammar, fluency) stabilise early with modest scale, whereas Content-oriented metrics (context, completeness, creativity) remain model size-sensitive, confirming the hierarchy outlined in Sec. **??**.

3.4 LLM-as-Judge Robustness

OpenAI Bias. GPT-4o is used for both data creation and evaluation, risking score inflation (Tab. 1).

Cross-Verification. Re-scoring with LLaMA-3.3-70B (Tab. 8) lowers all absolute values yet preserves the trends in Sec. 3.3; notably, LLaMA shows linear gains up to 157M, whereas GPT-4o plateaus at 54M.

3.5 Human Evaluations

3.5.1 Methodology

We recruited 100+ native speakers who produced 650+ ratings via an 8–10 min Google Forms survey (see Sec. C.4). Each rater saw the *exact* GPT-4o prompt and scored Grammar, Fluency, Completeness, Context, and Creativity on a 10-point Likert scale. Clear metric definitions, one cal-

ibrated example, and attention checks mitigated bias and fatigue (see C).

3.5.2 Results

Human–GPT-4o Delta Humans assign markedly lower scores than GPT-4o, aligning more closely with LLaMA:

Language	5M	54M	157M
Hindi	-1.63	-0.86	-0.90
Marathi	-4.03	-2.79	-1.87
Bangla	-2.15	-1.72	-1.88

Trend Replication Despite lower absolute scores, human raters reproduce key patterns from GPT-4o: early mastery of form (Grammar, Fluency) and persistent difficulty with content (Context, Completeness), confirming the validity of the Regional-TinyStories benchmark.

3.5.3 Statistical Validation

Z-score normalisation (C.3.3) followed by Kruskal–Wallis tests verifies intra- and inter-lingual differences at $p < 0.001$ (Tab. 16), supporting the robustness of our findings.

3.5.4 Inter-Rater Agreement

Krippendorff's α shows strong agreement for Hindi ($\alpha = 0.67$) and Bangla ($\alpha = 0.70$), but weak consistency for Marathi ($\alpha = 0.33$; Tab. 18), likely due to heterogeneous fluency among evaluators. Enhancing Marathi evaluation quality is a priority for future work.

3.5.5 Conclusion

Human evaluation corroborates the hierarchical emergence of linguistic capabilities observed with automated judges while revealing a systematic score inflation by GPT-4o. The consistent form-over-content gap across languages validates our methodology and highlights the necessity of human oversight. Future work will focus on improving evaluator calibration in Marathi and expanding participant diversity to further strengthen cross-lingual reliability.

3.6 Factual Prompt Evaluation

Context–Fact Correlation Manual analysis in Sec. B.3, anchored by the Hindi exemplar (Tab. 2), shows a strong positive correlation between precise context tracking and factual accuracy ($r > 0.8$; cf. Tab. A).

Scale Effects Larger SLMs excel on factual Q/A, while smaller ones lag yet display non-trivial factual competence, indicating generalisation beyond rote memorisation (Sec. H).

3.7 SLM *Language-Learning* Analysis

3.7.1 Limitations in Inter-Language Analysis

In line with Section H.1, 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.

Definition and Evaluation Priority

Language-learning is an SLM's ability to generate stories that are *context-aware*, *complete*, *creative*, *fluent*, and *grammatically correct* when trained on data in a target language. Following Sec. H.5, we treat **Context Awareness**—adherence of middle and end to the input prompt—as the primary evaluation metric; after all, a narrative that violates the prompt is unusable, irrespective of other qualities.

Methodology

For each language (Hindi, Marathi, Bangla) we train a 54M-parameter SLM on 2M synthetic *TinyStories*, sample 3,000 outputs, and score them with GPT-4o (Sec. B.1.2). Pairwise Mann–Whitney tests confirm inter-language differences ($p < 0.001$; Sec. H.7). Key robustness checks include (i) cross-evaluation with LLaMA-3.3-70B and (ii) the human study of Sec. C.

Findings

- **Context–Score Gap** All languages show lower means and higher variance for Context than for Grammar/Fluency, with Marathi displaying the widest gap (Grammar–Context $\Delta = 1.47$).
- **Consistency** Mean–SD correlations are strongly negative (Hindi/Bangla $r = -0.70$, Marathi $r = -0.77$): higher scores coincide with greater stability.
- **Support from Human Evaluations** Although inter-rater agreement on Marathi stories was lower—potentially weakening this finding—human judges consistently ranked Marathi's contextual performance as the lowest. In fact, the mean scores for both Hindi

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*

Table 2: Additional examples across languages in Sec. B

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

Tokenizer Name	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
Hindi							
Sarvam-1	0.518	7.734	7.783	7.806	8.554	8.912	8.158
SUTRA-mlt-256-v2	0.522	7.548	7.449	7.584	8.292	8.875	7.950
Tikoken	0.149	6.974	7.106	7.360	7.889	8.681	7.602

Table 4: **Tokenizer performance** for Hindi SLMs
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

Figure 3: Comparison of **54M param. SLM trained on Translated vs Synthetic Hindi Data** — Lighter is better

and Bangla lie above the entire 95% confidence interval for Marathi (see Tabs. **??**).

- **Ranking by Context** When Context takes precedence, *language-learning* in 54M models ranks: **Hindi** > **Bangla** > **Marathi**. This is trend supported by SOTA LLMs consistently performing poorly on Contextual Awareness for SLMs trained on Marathi *TinyStories* datasets (see Tab 1).

Conclusion

The Regional-TinyStories benchmark enables rapid, end-to-end training of distinct SLMs and LLM-as-a-Judge evaluation. Prioritising Context Awareness provides a principled, cross-lingual yardstick that reliably identifies the language in which an SLM achieves the most robust generative quality—currently Hindi—while highlighting avenues (e.g. Marathi) that require targeted data

or modelling improvements.

3.8 Tokenizer Comparison

Method

For each language we train a 54M SLM with Sarvam-1, SUTRA, and Tiktoken, then score outputs via LLM-as-Judge.

Results

Indic tokenisers outperform Tiktoken across all metrics; Sarvam-1 narrowly tops SUTRA despite a $\times 3$ smaller vocab and faster encoding.

Takeaway.

Regional-TinyStories offers a fast, context-first yardstick for selecting tokenisers in low-resource settings.

3.9 Effects of Training on Translated vs Synthetic Datasets

Rationale

Directly training and benchmarking separate LLMs for each dataset variant is prohibitively slow and costly. The **Regional-TinyStories** framework replaces this with ultra-compact SLMs per dataset as faithful surrogates, enabling rapid, low-cost comparisons (details in Sec. G.2).

Findings

Utilising the Regional-TinyStories framework, we train SLMs on both synthetically generated and translated datasets, observing across Hindi, Marathi, and Bangla, *synthetic-trained* SLMs outperform *translated-trained* counterparts on all five metrics (Tab. 19).

Conclusion

Regional-TinyStories delivers an end-to-end, data-agnostic pipeline for fine-grained evaluation beyond traditional metrics like LaBSE (Sec. G.1): train distinct SLMs, evaluate with an LLM judge, and select the dataset yielding superior context-first inference. This procedure generalises beyond synthetic vs. translated corpora, offering a scalable blueprint for benchmarking any pair of training datasets.

4 Conclusion

We generalise TinyStories beyond English by releasing 2 M–story corpora in Hindi, Marathi, and Bangla and training 5–157 M-parameter Small Language Models (SLMs) that attain over 90% of GPT-4o performance on key metrics despite orders-of-magnitude fewer parameters:contentReference[oaicite:5]index=5. The proposed **Regional-TinyStories** framework positions SLMs as time- and cost-efficient surrogates for Large Language Models, enabling rapid, variable-wise inference analysis and revealing an inflection point at 54 M parameters together with a stable Grammar $\rightarrow$ Fluency $\rightarrow$ Completeness $\rightarrow$ Context hierarchy:contentReference[oaicite:6]index=6.

A fine-grained tokenizer study confirms the Indic-specific Sarvam-1 outperforms SUTRA and Tiktoken, directly translating to higher context and fluency scores:contentReference[oaicite:7]index=7. Dataset comparisons show that synthetic corpora deliver up to 30% gains in context awareness over machine-translated data across all three languages, while grammar remains largely unaffected:contentReference[oaicite:8]index=8.

Results are cross-validated with an LLM-as-Judge ensemble (GPT-4o, LLaMA-3.3-70B) and a 100+ participant human evaluation, which reproduces the form-over-context gap and highlights modest score inflation in automated assessments:contentReference[oaicite:9]index=9.

Collectively, these findings establish Regional-TinyStories as a lightweight, reproducible pipeline for benchmarking tokenizers, datasets, and modelling choices in low-resource settings, and underscore the centrality of context awareness for evaluating language-learning in compact models.

Limitations

- **Unknown bias in SOTA LLMs & Tokenizers:** In line with Section H.1, 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

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

Impact Statement

Opportunities and Challenges

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

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

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

To ensure responsible use, we recommend:

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

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

Code and model repository

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

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

References

Catherine Arnett. 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. 2024a. Improving language models trained on translated data with continual pretraining and dictionary learning analysis. *Preprint*, arXiv:2405.14277.

Sabri Boughorbel, Md. Rizwan Parvez, and Majd Hawasly. 2024b. 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.

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.

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

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.

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

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

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.

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

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

Andrej Karpathy. 2022. nanogpt.

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

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

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

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent abilities in reduced-scale generative language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1242–1257, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. Tiktoken.

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

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

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.

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

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

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

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*. Open-source release of NLLB-200 and later variants.

Chris van der Lee and et al. 2019. Best practices for the human evaluation of automatically generated text. In *Proc. ACL 2019*.

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.

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

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

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

Appendices

Appendix Index

Complete Hyperparameter and Tokenizer Comparison Tables 14

Inference Evaluation and Factual Comparison 16

Human Evaluation of SLM Generated Stories 22

SLM Architecture, Training & Tokenizer Details 29

Synthetic Dataset Generation through LLM Prompting 32

Analysis of Synthetic Training Data Linguistic Diversity and Evaluation 34

Analysis of Translated Datasets & Synthetic vs. Translated Datasets 38

Statistical Analysis of Inference Results & SLM *Langauge-learning* analysis 42

Comments on Tokenizers and Renyi Entropy 47

Additional Benchmarking of SLM vs GPT-4 stories 50

A Complete Hyperparameter and Tokenizer Comparison Tables

A.1 Complete Hindi, Marathi & Bangla Hyperparameter Comparisons

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

Table 5: Models trained on **Hindi** TinyStories dataset - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

Hidden Size	Layer	Model Size	Eval Loss	Context	Completeness	Creativity	Fluency	Grammar	Overall
64	**2**	4.46	2.680	6.046	7.119	7.292	7.607	8.332	7.279
64	**6**	4.65	2.343	6.850	7.430	7.389	8.036	8.809	7.703
64	**12**	5.00	2.224	6.050	7.520	7.465	8.094	8.813	7.588
512	**2**	41.00	1.773	6.474	8.005	7.801	8.726	8.980	7.997
512	**6**	54.00	1.662	7.154	8.127	7.902	8.854	9.146	8.236
512	**12**	73.00	1.654	6.992	8.003	7.814	8.757	9.096	8.132
1024	**2**	94.00	1.682	7.290	7.689	7.199	8.193	8.855	7.845
1024	**7**	157.00	1.640	7.641	7.697	7.200	8.257	8.940	7.947
GPT-4o-mini	**-**	-	-	8.630	9.459	8.416	9.359	9.441	9.061
GPT-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

Table 6: **Marathi** - Hyper-parameter Comparison — Lighter the better — 3000 stories per configuration evaluated

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

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

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 8: **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***

A.3 Complete Tokenizer Evaluation Comparison

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

Table 9: **Tokenizer performance** across Hindi-Marathi-Bengali
Sarvam Tokenizer — 54M (E=512, L=6) Model — Lighter is better

B Inference Evaluation and Factual Comparison

B.1 SLM Inference and Evaluation Methodology

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

B.1.1 Preparing Prompts for SLM Evaluation

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)[5]. 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

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

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 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": "...",
"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. 10, 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. 12, 13, and 14. This behaviour is further statistically analysed and supported in Sec. H.6, where Marathi models of all sizes exhibit similar difficulties in preserving prompt relevance compared to their Hindi and Bangla counterparts.

B.3 Factual Prompt Evaluation

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

Using the multilingual factual benchmark in Tab. 11, 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.

[5] 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 10: **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 11: 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 12: **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 13: **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 14: **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 and 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 children-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 $\pm$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 15: 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 16: 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 17: **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	Good	0.748	0.662	0.733	0.719	0.663
Hindi	0.666	Tentative	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 18: **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. 15.

C.3.2 Comparing Human Evaluations to GPT-4o Evaluations

C.3.3 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:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad (1)$$

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.3.4 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

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 19: **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
Δ vs. GPT-4o	-4.75	-4.67	-2.41	-3.67	-4.64	-4.03
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]
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 20: **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 21: **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

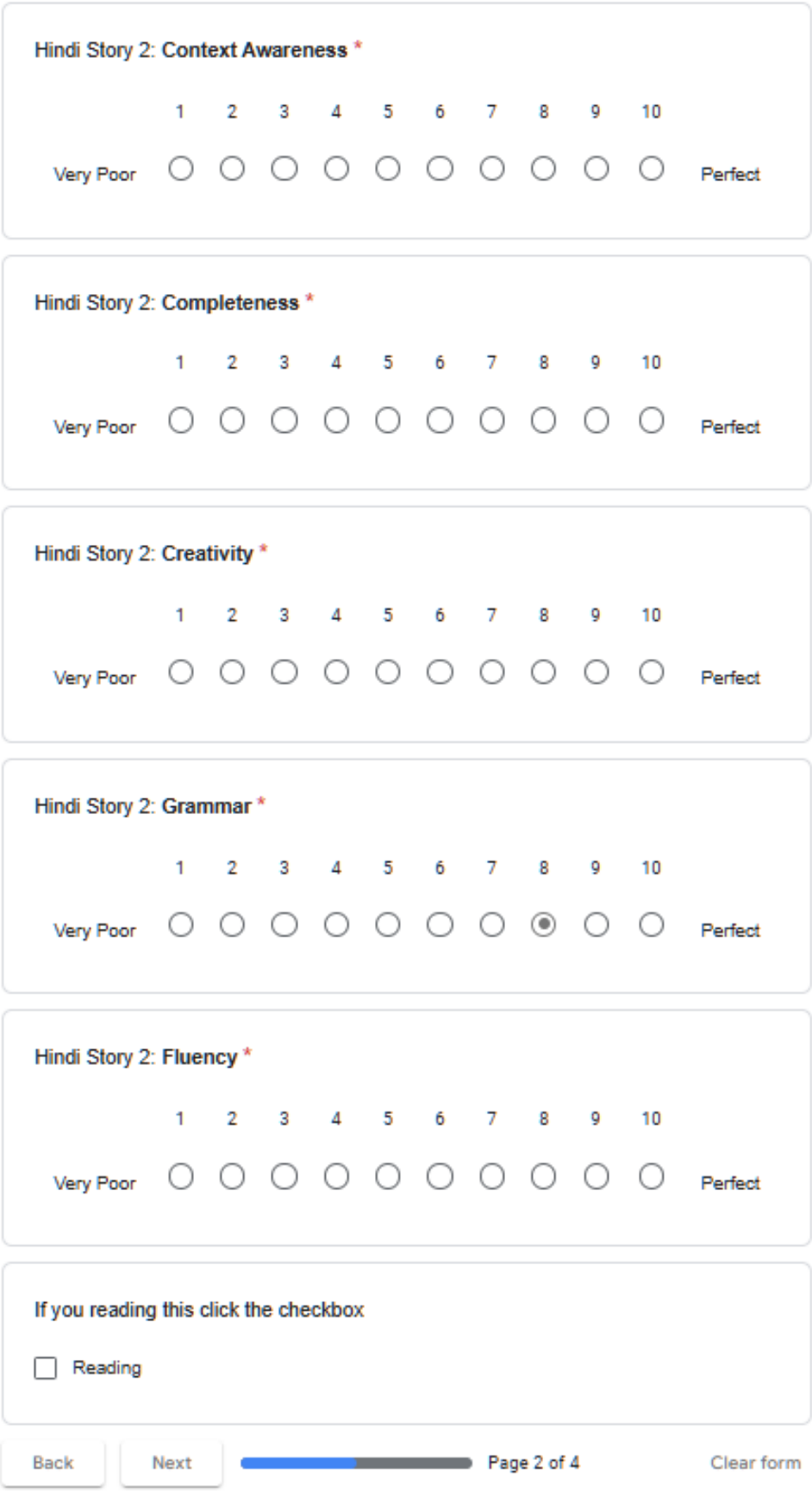


Figure 7: **Sample story to evaluate (Part-II)** in one of our Hindi Google Forms

D SLM Architecture Training & Tokenizer Details

The implementation for this section is available at `training-inference/` in our repository[6].

D.1 nanoGPT based SLM Architecture

We build all Regional-TinyStories models on **nanoGPT** (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:

$$y = x + \text{MHA}(x), \quad (2)$$

$$z = y + \text{MLP}(\text{LN}(y)), \quad (3)$$

$$\text{Block}(x) = \text{LN}(z). \quad (4)$$

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 22: **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.*

[6] Link To: Code & Datasets

D.3 Training (WANDB) Logs

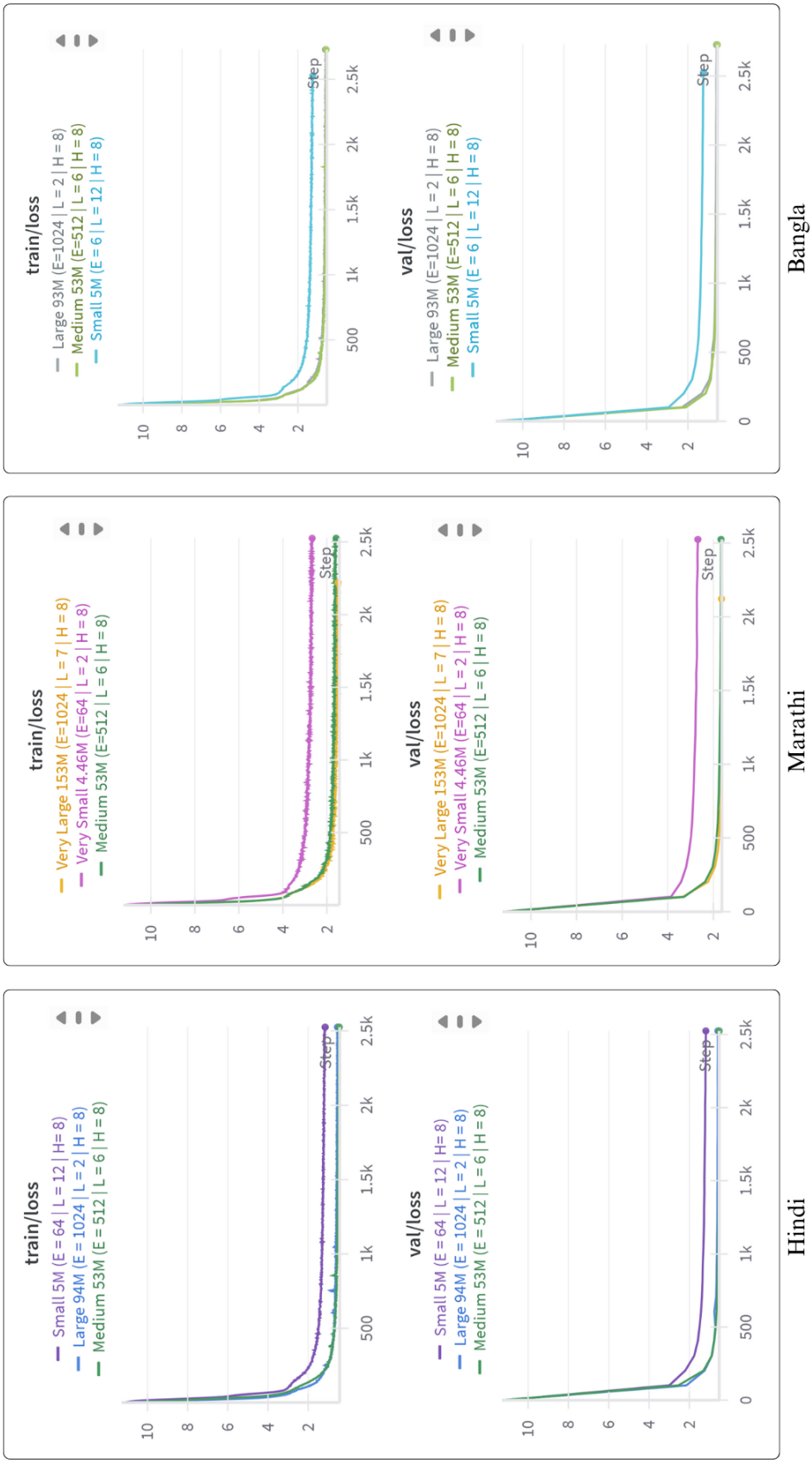


Figure 8: Training & Testing loss curves for Small, Medium and Large Models across Hindi, Marathi & Bangla

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 23: 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 Analysis: Sarvam vs. SUTRA vs. 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).

Tokenizer Evaluation and Selection. In our experiments, Sarvam attains the lowest evaluation loss among the two Indic tokenizers and exhibits the highest fluency as judged by LLM-as-a-judge scores. This suggests that its compact yet Indic-aware vocabulary better captures morphological boundaries across diverse languages. SUTRA follows closely; its larger vocabulary offers slight benefits in performance but at the cost of increased computation. By contrast, Tiktoken achieves lower perplexity—likely due to over-segmentation—yet results in notably weaker narrative quality, reinforcing prior concerns that English-centric BPE introduces cross-language unfairness.

Sarvam as the Preferred Tokenizer. As detailed in Table 9, Sarvam-1 outperforms SUTRA-mlt-256-v2 across all evaluation metrics. In addition to superior performance, Sarvam offers faster tokenisation and maintains a significantly smaller vocabulary size. This compactness translates to models with substantially fewer total parameters, owing to reduced Token-embedding and Output-projection matrix sizes. Given this balance of efficiency and effectiveness, we adopt Sarvam-1 as our default tokenizer for all analyses unless otherwise specified.

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)[7].

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.

[7]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

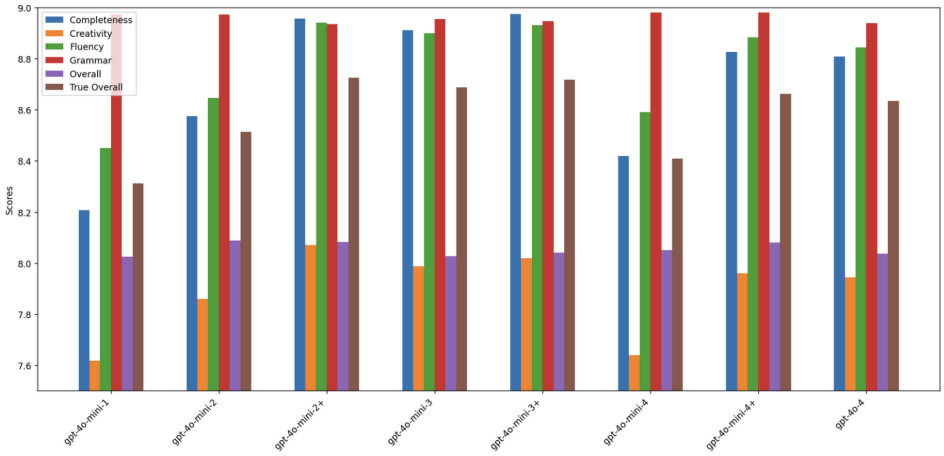


Figure 9: **Prompt Complexity Comparison**
gpt-4o-mini-1: Model used = GPT-4o-mini, Prompt complexity = 1
Prompt complexity = 2+ (where + indicates increased word limit of 350–500 words)

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[8].

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 24: Average ROUGE F1 Scores (English)

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

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

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 25: English TinyStories ROUGE scores sample

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

Table 26: 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.

[8] 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 pre-trained 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.

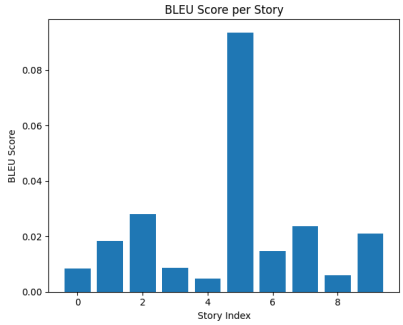


Figure 11: **BLEU** for 10 random Bangla stories.

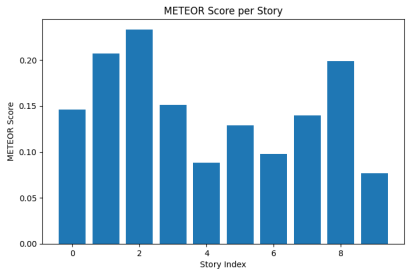


Figure 12: **METOR** for 10 random Bangla stories.

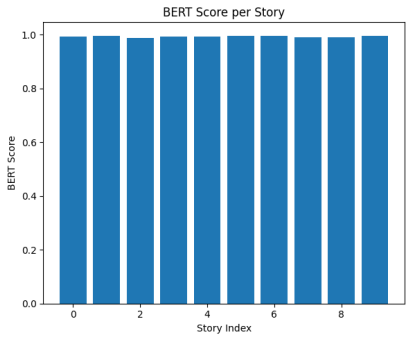


Figure 13: **BERT** for 10 random Bangla stories.

F.2.3 Co-relation Analysis

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

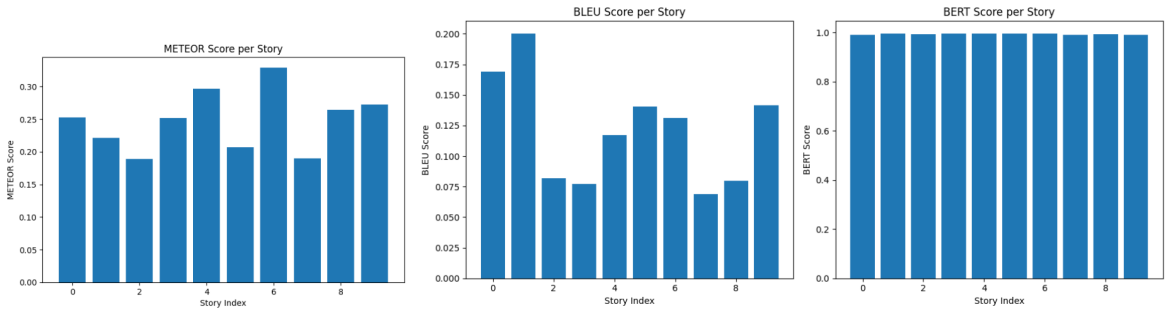


Figure 14: BLEU, BERT & METOR Scores for 10 random **Hindi** stories.

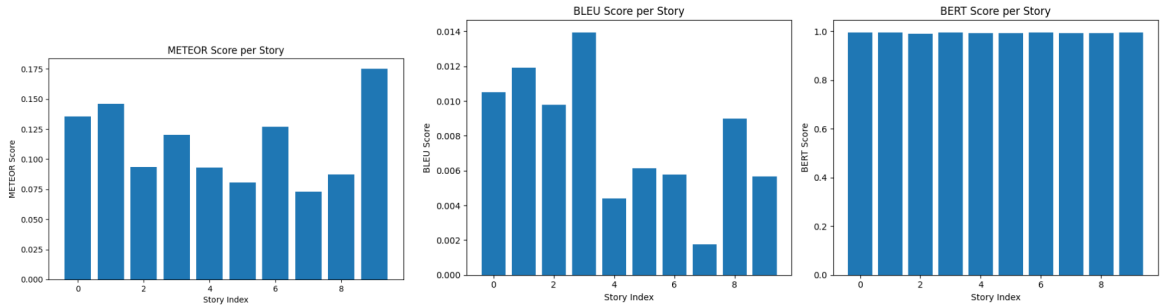


Figure 15: BLEU, BERT & METOR Scores for 10 random **Marathi** stories.

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

Table 27: 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 & Synthetic vs. 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 (next Sec. G.2). 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

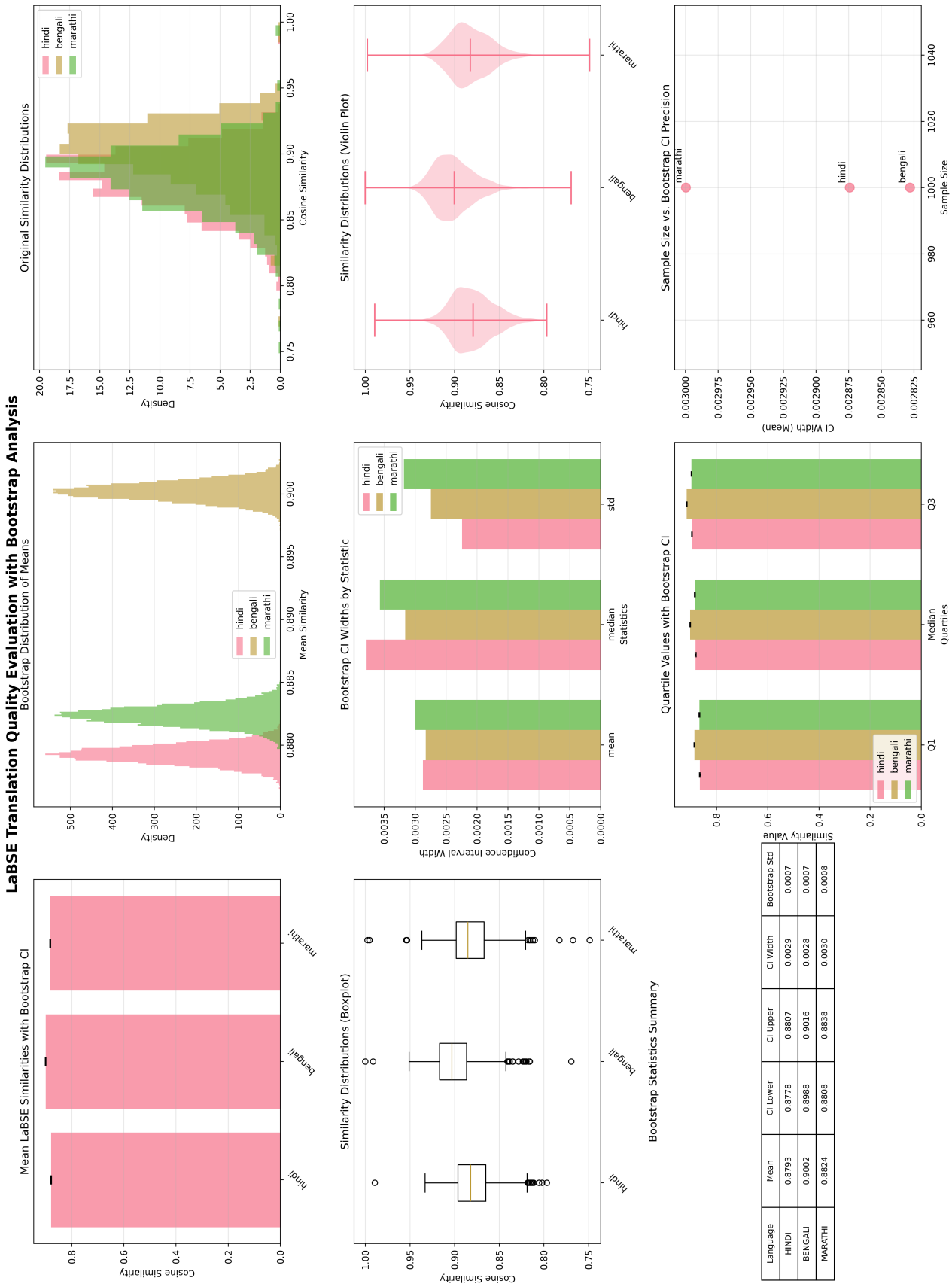
Our analysis establishes LaBSE evaluation as a valuable analysis step for translated datasets. The strong negative correlation between translation quality variance (measured by IQR and standard deviation) and model performance suggests that consistency in translation quality may be as important as mean 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. (2024b)

This multi-faceted evaluation approach—combining LaBSE semantic similarity with downstream inference metrics by utilising our Regional-TinyStories framework for dataset comparisons (next Sec. G.2)—provides a comprehensive framework for assessing and selecting training data for low-resource language modelling.



Bootstrap Statistics Summary

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

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

Figure 19: Performance comparison of **Models trained on Translated vs Synthetic Data** performance across Hindi-Marathi-Bangla. *54M (E=512, L=6) Model — Lighter is better*

G.2 Regional-TinyStories Framework for Comparing Synthetic and Translated Training Data

G.2.1 Motivation

The motivation for this section follows that outlined in the Introduction of the main paper–training distinct LLMs for each variation (here datasets) to directly measure inference quality is severely cost and time-inefficient, we circumvent this via our Regional-TinyStories framework, which utilises ultra-compact TinyStories-based SLMs as a time and cost-efficient proxy for LLMs.

G.2.2 Methodology

Our objective is to identify which training dataset yields the highest inference quality for Small Language Models (SLMs), evaluated across five dimensions: Context Awareness, Creativity, Completeness, Fluency, and Grammar. Using the Regional-TinyStories framework, we train one SLM from scratch per dataset and generate 3,000 stories from 1,000 predefined prompts (See Sec. B.1). These outputs are then evaluated by GPT-4o, serving as an LLM-as-a-Judge. We rank datasets via inference scores of corresponding SLMs, based primarily on context awareness, followed by the overall score.

G.2.3 Comparing Synthetically generated vs Translated Datasets

While Sec. G.1 quantifies semantic loss accompanying translation using LaBSE, our Regional-TinyStories framework enables a more fine-grained, inference-based evaluation of training SLMs on synthetically generated versus translated data.

As shown in Tab. 19, SLMs trained on synthetic data consistently outperform those trained on translated data across all three languages and all five evaluation metrics. The most pronounced gap appears in Context Awareness—the key metric for assessing a model's ability to follow instructions, followed by notable declines in Completeness and Creativity. In contrast, Grammar remains largely unaffected, indicating that while translation introduces semantic noise, surface-level linguistic structure is largely preserved.

These findings highlight the strength of our Regional-TinyStories framework in uncovering subtle performance degradations that conventional metrics, such as LaBSE alone, cannot capture. Specifically, translation-based training disproportionately impacts a model's ability to retain contextual relevance, even as grammatical correctness is maintained.

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[9].

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 Tabs. 5, 6, and 7) 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.

SLM *Language-learning* Analysis We analyse and quantify the effects on short story inference of training SLMs across Hindi, Marathi and Bangla synthetic TinyStories dataset.

[9]Link To: Code & Datasets

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

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.

We interestingly uncover that the **157M** (E: 1024, L: 7) 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

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 28: **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 29: **Standard Deviation Comparison** Across Metrics and Languages—*for 54M (E:512, L:6) models trained on Hindi, Marathi and Bangla synthetic TinyStories datasets*

metrics with lower means (like context awareness) show higher standard deviations. *This suggests that as performance on a particular aspect improves, consistency also increases.*

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

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

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 (0.42, 0.60) than those related to content (context, completeness, creativity) (1.18, 0.92, 0.62) (see Tab. 29).

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. 30) *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. 30), 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—SLM Language-learning analysis

H.5 Context–Is–Key

We define *Context Awareness as the degree to which a story, especially its middle and end, adheres to its input prompt.* This metric is paramount: a narrative may be complete, creative, fluent, and grammatically correct, but it remains irrelevant if it fails to follow the user's instructions. Additionally, as demonstrated in Sec. H.4, contextual reasoning remains the most difficult to develop, an emergent metric with a strong sensitivity to increasing model size.

These results underscore *Context Awareness* as the foremost metric for evaluating SLM inference quality and for enabling robust cross-lingual comparisons (see next section) among models trained on different language datasets.

H.6 Inter Language Performance Relationship & SLM *Langauge-learning* analysis

Overview Building on our characterisation of metric hierarchies and the central role of Context Awareness, we now perform a cross-lingual

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

Table 31: **Overall Performance Statistics** by Language—*for 54M param models trained on Hindi, Marathi and Bangla synthetic TinyStories datasets*

comparison using our Regional-TinyStories framework, comparing the *language-learning* abilities of our SLMs across Hindi, Marathi, and Bangla.

Language Learning We define *language-learning* as each SLM's ability to generate stories that are context-aware (with respect to input prompt), complete (corresponding to logical endings with moral values), creative, fluent, and grammatically correct when trained on datasets in the language under consideration. *For example, a 54M model's Hindi* language-learning *refers to the quality of stories it can generate when trained on our synthetically generated Hindi TinyStories dataset.*

Methodology For each language (using our corresponding synthetic TinyStories datasets), we train a 54M-parameter SLM on our 2M short stories datasets and generate 3,000 outputs per model. GPT-4o then evaluates (see Sec. B.1.2) these outputs (see Sec. A), and Tab. 28 (above) presents the resulting scores. These evaluation scores are then used to make downstream comparisons, in this case ranking the aforementioned *language-learning* capabilities with respect to the highest scoring SLMs.

Statistical Significance Analysis To highlight the robustness of our inter-language comparisons, we detail the statistical significance of all inter-language pairs in the next section (Sec . H.7).

H.6.1 Language-Specific Co-relations

All three languages show a moderate to strong negative correlation between means and standard deviations, with Marathi having the strongest inverse relationship (r = - 0.77), while both Bangla and Hindi show identical correlations (r = - 0.70). This negative correlation suggests more consistent performance in languages where the models score higher, i.e. Hindi and Bangla.

H.6.2 Language-Specific Consistency Patterns

Marathi shows higher standard deviations across all metrics (average: 0.87) compared to Hindi (0.65) and Bangla (0.69 (Tab. 29), suggesting greater variability in performance. This is particularly evident in context awareness (Marathi: 1.42 vs. Hindi: 1.01) and overall scores (Marathi: 0.67 vs. Hindi: 0.52).

H.6.3 Form-Content Divide

This divide, between form-oriented capabilities (grammar and fluency) and content-oriented skills (context awareness, completeness, creativity), is most pronounced in Marathi and least evident in Hindi.

H.6.4 Language-Specific Gap Patterns

Marathi shows notably larger evaluation score gaps between grammar and other metrics (In Tab. 30 Grammar−Context = 1.47, Grammar−Completeness=1.31) compared to Hindi and Bangla. This suggests that while Marathi models achieve reasonable grammar scores, they struggle more with contextual coherence and narrative completeness compared to models in other languages.

H.6.5 Conclusion–SLM *Langauge-learning* analysis

While the 54M Marathi SLM achieves the highest overall score (unweighted average across Context, Completeness, Creativity, Grammar, and Fluency), its Context Awareness score is substantially lower than its Hindi and Bangla counterparts.

Context Awareness, identified as our primary evaluation metric, requires a sensitive, fine-grained comparison to prioritise it, either by considering it first–amongst other metrics, or by applying a weighted average that assigns greater importance to Context Awareness.

Under this more rigorous rubric–ranking by *Context Awareness* first, the *language-learning* capabilities of our 54M-parameter SLMs rank as follows: Hindi > Bengali > Marathi.

Limitations Our *language-learning* rankings are constrained by ambiguities in GPT-4o's cross-lingual evaluation methodology and potential proficiency differences across Hindi, Marathi, and Bangla (see Sec. H.1). To address these constraints, we complement our quantitative

results with the human evaluation study detailed in Sec. C.

Next sections—Statistical Significance Analysis

H.7 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.7.1 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.7.2 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.7.3 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.7.4 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.7.5 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 Comments on Tokenizers and Rényi Entropy

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

In our comparative analysis of tokenizer performance, we sought to control for variations in dataset size across different languages. Following the methodology proposed by (Arnett and Bergen, 2025), we implemented byte premium scaling. (Arnett and Bergen, 2025) argue that disparities in language model performance across languages are often not due to the inherent morphological complexity of the languages themselves, but rather to inequities in the size of the training data. The "byte-premium" refers to the concept that different languages and their orthographies have varying encoding efficiencies; the same semantic information may require a different number of bytes to be stored depending on the language.

To ensure a fair comparison, it is therefore crucial to normalize the datasets by their size in bytes, rather than by word or document count. This approach ensures that each language is represented by a corpus of equivalent information content, from an encoding perspective. Our implementation of byte premium scaling is detailed below:

Our implementation pipeline for byte premium scaling and subsequent analysis involved several steps, each carried out by a dedicated script. The overall workflow is outlined below.

1. **Data Extraction**:
   - Script: `extract_raw.py`
   - Description: Load the `train` split of each Hugging Face dataset (Hindi, Marathi, Bengali), extract the "story" field, replace newline characters with spaces, and write each cleaned story as a single line in an output `.txt` file.
2. **Byte-Level Downsampling**:
   - Script: `sample_to_target_bytes.py`
   - Description: Shuffle all lines of a given raw text file (fixed random seed for reproducibility), then sequentially add lines to a new file until a specified target byte-size as outlined in `byte-premium-tool` for Hindi, Marathi and Bengali is reached. This produces scaled corpora in bytes that are identical across languages.

Table 32: Mean Rényi entropy, variance, and 95% confidence intervals by language and tokenizer.

Language	Tokenizer	Mean Entropy	Variance ($\times 10^{-7}$)	95% CI Lower	95% CI Upper
Hindi	SUTRA	7.101	0.915	7.100	7.101
Hindi	Sarvam	6.280	1.400	6.279	6.280
Marathi	SUTRA	7.678	2.860	7.677	7.679
Marathi	Sarvam	6.506	0.539	6.505	6.506
Bengali	SUTRA	7.319	1.110	7.319	7.320
Bengali	Sarvam	6.304	0.733	6.303	6.304

3. **Tokenization**:
   - Script: `prepare_tokenized_corpora.py`
   - Description: For each language–tokenizer pair ("sutra" and "sarvam"), read the scaled raw text and output a tokenized file where each line contains space-separated tokens corresponding to the input text.
4. **Corpus Splitting for Statistical Analysis**:
   - Script: `split_scaled_corpora.py`
   - Description: Shuffle the lines of each tokenized file, then split it into five equal-sized, non-overlapping subsets. These splits allow computation of the mean and variance of the Rényi entropy metric across folds.

I.2 Tokenization Efficiency via Rényi Entropy

To quantitatively compare the tokenization efficiency for each language, we calculated the normalized Rényi entropy. As noted by (Arnett and Bergen, 2025), Rényi entropy is a useful metric for tokenization quality as it measures the evenness of the token frequency distribution. We used the `compute_stats.py` script to perform this analysis.

The script calculates the Rényi entropy (with $\alpha = 2.5$) for each of the five splits for every language–tokenizer pair. It then computes the mean Rényi entropy, variance, and a 95% confidence interval. The results are summarized in Table 32.

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

To determine if the observed differences in mean Rényi entropy between the languages were statistically significant, we performed pairwise Welch's t-tests. This test is appropriate as it does not assume equal variances between the samples. The `compute_stats.py` script performs these tests. The results of the t-tests are presented in Table 33.

Table 33: Pairwise Welch's t-test results for Rényi entropy differences.

Tokenizer	Pair	t-statistic	p-value
SUTRA	Hindi–Marathi	−2.100	<5.000
SUTRA	Hindi–Bengali	−1.086	<5.000
SUTRA	Marathi–Bengali	1.273	<5.000
Sarvam	Hindi–Marathi	−1.146	<5.000
Sarvam	Hindi–Bengali	−1.153	<5.000
Sarvam	Marathi–Bengali	1.266	<5.000

The extremely small p-values (all $p \ll 0.001$) across all pairwise comparisons indicate that the differences in Rényi entropy scores between Hindi, Marathi, and Bengali are statistically significant for both the `sutra` and `sarvam` tokenizers. This robust statistical evidence allows us to confidently conclude that the tokenizers exhibit different levels of efficiency for each of the languages in our study, even after controlling for dataset size via byte-premium scaling. The very large absolute values of the t-statistics further underscore the magnitude of these differences. This detailed analysis provides a solid foundation for our main paper's discussion on tokenizer performance.

1. Table 34 summarizes the Rényi entropy for Hindi, Bengali, and Marathi using the Sarvam-1 and SUTRA tokenizers. The vocabulary sizes for these tokenizers are approximately 68K for Sarvam-1 and 256K for SUTRA, which influences the entropy values and reflects their distinct tokenization approaches.

2. A primary observation is the consistently high entropy for Marathi, particularly at lower α values. At α=0.5, which emphasizes rare tokens, Marathi tokenized with Sarvam achieves an entropy of 10.926, the highest across all configurations. This quantitatively supports the assertion that Marathi's agglutinative morphology creates a significantly long tail of infrequent subwords, a known challenge in modeling (Dolamic and Savoy, 2010; Arnett and Bergen, 2025). Sarvam-1 balances vocabulary size and entropy, capturing meaningful words efficiently. SUTRA's larger vocabulary and generally high entropy reflect a finer-grained and more complex token distribution.

3. The entropy trends at different α values further elucidate the characteristics of these tokenizers:

   (a) $\alpha = 0.5$ (emphasizing rare tokens): Sarvam consistently exhibits higher entropy than SUTRA for all three languages. This indicates that Sarvam's tokenization is more sensitive to the diversity of rare tokens, effectively capturing a more complex distribution across the vocabulary's long tail.

   (b) For ≥1.0 (or ≥1.5 for Marathi), the trend reverses, and SUTRA consistently yields higher entropy. For instance, at Shannon Entropy (α=1.0), SUTRA's entropy for Bangla is 8.745 compared to Sarvam's 8.649. This gap widens at higher α values that focus on common tokens. At α=2.5, SUTRA's entropy for Marathi is 7.670, substantially higher than Sarvam's 6.506.

   This crossover reveals a fundamental trade-off. Sarvam's tokenization appears more efficient for representing the diversity of rare morphological variants. In contrast, SUTRA's much larger vocabulary provides a finer-grained and more complex representation of mid-frequency and common subword patterns, resulting in a more diverse distribution for the head and body of the token frequency list. This suggests that while Sarvam may be robust for rare phenomena, SUTRA can better capture the nuanced diversity within common language usage.

4. According to Arnett and Bergen (2025), agglutinative languages tend to have higher Rényi entropy compared to fusional languages. A study comparing Hindi, Marathi, and Bangla notes that Marathi's agglutinative structure creates more complex inflectional patterns, which require distinct stemming strategies for information retrieval tasks. In contrast, Bangla's simpler fusional morphology differs from Marathi's suffix-heavy word formation (Dolamic and Savoy, 2010). The elevated entropy values for Marathi, especially at $\alpha = 0.5$, underscore its agglutinative complexity and the long tail of rare subwords, confirming that tokenizers with larger, more tailored vocabularies are better suited to capture the nuances of Indic morphology.

Table 34: Rényi entropy (H_α) of Hindi, Marathi, and Bangla with Sarvam and SUTRA tokenizers.

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

J Additional Benchmarking of SLM vs GPT-4 stories

J.1 COMET-DA and LaBSE

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

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

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

J.2 Final Conclusions

The following observations are made based on the results in Tables. 35 to 40:

- Hindi models show steady improvement in COMET-DA scores as parameters increase, with peak performance at 95M parameters before plateauing. Similarly, LaBSE cosine similarity improves with model size, reaching optimal performance at 54M and 95M checkpoints.
- Marathi models demonstrate an upward trend in COMET-DA scores but with notable fluctuations at larger sizes (95M and 157M). LaBSE similarity is highest at 54M model size.
- Bangla models consistently outperform both Hindi and Marathi, achieving higher COMET-DA scores that suggest better semantic alignment with GPT-4o outputs. Their LaBSE cosine similarity remains consistently high across all model sizes, indicating robust semantic similarity and stronger language modeling capabilities under the evaluated conditions.
- Across all languages, larger models generally achieve higher scores on both metrics, though performance plateaus or fluctuates at very large sizes (95M-157M). Bangla models consistently outperform the others, highlighting their superior semantic alignment with GPT-4o outputs, based on this metric.

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

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

Table 35: 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 36: 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 37: 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 38: 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 39: 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 40: LaBSE similarity for Bangla checkpoints (ascending size).