BhashaKritika: Building Synthetic Pretraining Data at Scale for Indic Languages

Guduru Manoj*, Neel Prabhanjan Rachamalla*, Ashish Kulkarni, Gautam Rajeev, Jay Piplodiya, Arul Menezes, Shaharukh Khan, Souvik Rana, Manya Sah, Chandra Khatri, and Shubham Agarwal

[1] Krutrim, India

{neel.rachamalla1, ashish.kulkarni, shubham.agarwal1}@olakrutrim.com

Abstract

In the context of pretraining of Large Language Models (LLMs), synthetic data has emerged as an alternative for generating high-quality pretraining data at scale. This is particularly beneficial in low-resource language settings where the benefits of recent LLMs have been unevenly distributed across languages. In this work, we present a systematic study on the generation and evaluation of synthetic multilingual pretraining data for Indic languages, where we construct a large-scale synthetic dataset *BhashaKritika*, comprising 540B tokens using 5 different techniques for 10 languages. We explore the impact of grounding generation in documents, personas, and topics. We analyze how language choice, both in the prompt instructions and document grounding, affects data quality, and we compare translations of English content with native generation in Indic languages. To support scalable and language-sensitive evaluation, we introduce a modular quality evaluation pipeline that integrates script and language detection, metadata consistency checks, n-gram repetition analysis, and perplexity-based filtering using KenLM models. Our framework enables robust quality control across diverse scripts and linguistic contexts. Empirical results through model runs reveal key trade-offs in generation strategies and highlight best practices for constructing effective multilingual corpora.

Extended version — https://arxiv.org/pdf/2511.10338

1 Introduction

Most state-of-the-art LLMs (Grattafiori et al. 2024; Abdin et al. 2025) are trained predominantly on English corpora, available in abundance, leaving many of the world's other languages underrepresented in both training data and model performance. emphasize the finite nature of available pretraining data, often sourced from CommonCrawl (Common Crawl 2007) and the need for alternative approaches to progress the state of LLMs. Even when multilingual datasets exist, they often suffer from issues related to quantity, quality, domain bias, diversity and inconsistent formatting (Conneau et al. 2020). Thus, open-access pretrained models with strong multilingual capabilities remain limited, especially for low-resource and morphologically rich Indian languages. Hindi, for instance, does not appear in the top 20 languages of Common Crawl despite being the third most spoken globally (Penedo et al. 2023) and Indian languages collectively constitute less than 1% (Kallappa et al. 2025). This scarcity of both data and models presents a major barrier to the development of culturally inclusive LLMs especially with recent data constrained scaling laws (Muennighoff et al. 2023) arguing that model performance show degradation after 4 epochs on repeated data.

Synthetic data generation has thus emerged as a viable approach where, training data is artificially generated while mirroring the features, structures, and statistical attributes of real-world data (Nadas, Diosan, and Tomescu 2025; Liu et al. 2024; Yu et al. 2023). This offers a compelling alternative to conventional web-scraping and manual curation while providing control and diversity compared to web data. By leveraging existing LLMs as generators, it is possible to create large-scale, language-diverse corpora that is customizable and replicable (Wang et al. 2022b; Taori et al. 2023; Longpre et al. 2023). The Phi series of models (Gunasekar et al. 2023; Abdin et al. 2024) focused on proprietary synthetic data as part of their pre-training corpus and showed its efficacy in their training pipeline. Ben Allal et al. (2024), created the open-source *Cosmopedia* consisting of 25B English synthetic tokens, grounded in web documents. Ge et al. (2024) introduced *PersonaHub*, a collection of English personas, that are then used for persona-grounded synthetic generation. Here, a 'persona' is defined as 'a person with specific professional experiences and cultural backgrounds having unique interests in reading and writing'.

In this work, we propose a pipeline for generating high-quality synthetic pretraining text data focusing on both Indian languages and local context. Our approach builds on previous work and involves language-aware prompt engineering, style and domain variation, and automated quality filtering to ensure broad linguistic coverage and coherence.

Specifically, we make the following contributions:

- We develop a modular pipeline for generating large-scale high-quality synthetic Indic multilingual corpora. We design prompt templates, data curation pipelines, generation strategies, and conduct ablations across languages and models that collectively ensure that the generated data is factually grounded, knowledge-dense, rich in Indic cultural context, and topically diverse.

*These authors contributed equally.

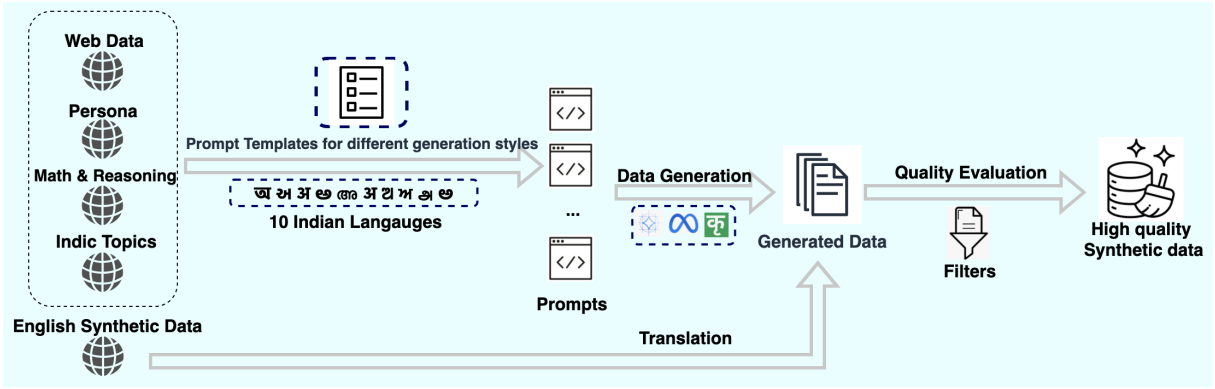

Figure 1: Overview of Synthetic data generation techniques (Section 3) followed by Quality Evaluation (Section 4). We follow 5 approaches across 10 Indian languages using a pool of Multilingual LLMs to generate a large scale *BhashaKritika* corpora.

- We propose a novel method for constructing math-focused pretraining data by transforming instruction-tuned datasets into pretraining-style corpora through controlled synthetic generation.
- We implement an automated quality filtering pipeline, covering language consistency, fluency, heuristic filters, statistical filters, quality classifiers, bias detection and mitigation strategy in the generated data.
- We use our synthetic generation pipeline to generate *BhashaKritika*, a 540B tokens high-quality Indic multilingual synthetic corpus. We also share a part of this data for public use[1].
- We perform extensive analysis with additional two controlled experiments including annealing as well as pretraining a 1B param from scratch and show models trained on synthetic data continue to improve and closely match the one trained with the real data.

2 Related Work

2.1 Web Crawled Datasets

For text-based pretraining, large-scale datasets such as The Pile (Gao et al. 2020), C4 (Raffel et al. 2020), RedPajama (Computer 2023), RefinedWeb (Penedo et al. 2023), Dolma (Soldaini et al. 2024), DataComp-LM (Li et al. 2024b), and FineWeb (Penedo et al. 2024a) have been instrumental in training LLMs. Nemotron-CC (Su et al. 2024) further refines this effort with a high-quality subset of 6.4T tokens while MegaMath (Zhou et al. 2025) filters for Math datasets. Mostly sourced from CommonCrawl (Common Crawl 2007), these datasets however, predominantly feature English and other high-resource languages, offering limited coverage of Indian languages or culturally grounded content. More recently, FineWeb2 (Penedo et al. 2024b) introduces broader language coverage, however only a small portion, around 40B words, pertains to Indian languages.

[1]https://huggingface.co/datasets/krutrim-ai-labs/BhashaKritika

2.2 Indic LLM Research

Most prior work has focused on adapting existing English-dominant models through fine-tuning or continued pretraining on Indic language corpora (Kohli et al. 2023; Gala et al. 2024; Sarvam 2023). In contrast, only a few models (Krutrim 2024; Sarvam 2024) have been trained from scratch, aiming to create more culturally inclusive LLMs for the Indian context. Alongside model development, there have been ongoing efforts to curate multilingual datasets focused on Indian languages. One of the earlier efforts in this direction, the IndicNLP Corpora (Kunchukuttan et al. 2020), compiled around 2.7B tokens for 10 Indic languages from web-based content which was later expanded to IndicCorp (Kakwani et al. 2020), comprising 8.8B tokens across 11 Indian languages and English. Gala et al. (2024) released Indic Instruct Data v0.1, a Hindi instruction-tuning dataset derived through translation of pre-existing instruction sets. Additionally, the Sangraha corpus (Khan et al. 2024) offers a collection of 251B tokens covering 22 Indian languages, nonetheless, its scale remains modest compared to the much larger corpora generally available for English (in $5-15$T tokens) and the other Western languages.

2.3 Synthetic Data Generation

Synthetic data techniques have become a valuable resource for enriching both fine-tuning and pretraining corpora. Early instruction-tuning methods include Self-Instruct (Wang et al. 2022a), Evol-Instruct (Xu et al. 2023) and Magpie (Xu et al. 2024) to name a few. Beyond fine-tuning, synthetic data for pre-training has also shown promise, notably in the proprietary Phi models (Li, Bubeck et al. 2023; Abdin et al. 2024). Open-source alternative Cosmopedia (Ben Allal et al. 2024) offer 25B tokens of diverse synthetic text generated in English. Recent work has also explored persona-based generation to increase diversity and alignment with PersonaHub introducing 1M synthetic personas (Ge et al. 2024) which Nemotron-Personas further aligns personas with demographic and psychological traits (Meyer and Corneil 2025). Similar techniques have been applied in multimodal settings (Yang et al. 2025) and domain-specific tasks like math and reasoning (Lambert et al. 2024).

Odumakinde et al. (2024) proposed a multilingual arbitrage framework to further improve teacher model selection across languages. Finally, synthetic data scaling laws proposed by (Qin et al. 2025) emphasize the interplay of quantity, diversity, and generation methods. We build upon these works to generate pre-training synthetic data for Indian languages.

3 Synthetic Data Generation Techniques

We develop a pipeline to generate synthetic data at scale and demonstrate how we used it to generate 540B tokens of high-quality Indic synthetic data. Our pipeline leverages different Web data sources (both direct and derived) as context, multiple generation techniques and output styles that together ensure that the generated synthetic data is factually grounded, knowledge-rich, and topically diverse.

3.1 Document Grounded Generation

Following work on English synthetic data generation (Ben Allal et al. 2024; Su et al. 2024; Li, Bubeck et al. 2023), we leverage multilingual LLMs, prompted with documents from the Web as context, to generate Indic synthetic data in different knowledge-rich formats and creative styles. (Ben Allal et al. 2024; Su et al. 2024; Maini et al. 2024; Li, Bubeck et al. 2023; Gunasekar et al. 2023). In addition to using documents from English FineWeb (Penedo et al. 2024a) and multilingual FineWeb2 (Penedo et al. 2024b) directly, we also selectively curate "Indic context" documents. The Indic context documents are identified using a FastText-based classifier (Joulin et al. 2016) trained on 93K annotated documents. We also perform clustering by adapting Huggingface text-clustering[2] to get broad topics for Indian context data, that we optionally append to the prompts for document grounded generations.

We evaluate (via human annotation) different multilingual LLMs based on their quality of direct generations in a language xx and generations in En followed by their translation to xx. Table 1 shows the language-model mapping that we followed for our synthetic generation. In addition to ensuring language-specific quality, the use of multiple LLMs in our pipeline also alleviates model-specific biases, avoids model collapse, and encourages generalization, diversity, and robustness in our generated data (Agarwal, Bozdag, and Hakkani-Tür 2025; Odumakinde et al. 2024).

Taking inspiration from prior works (Ben Allal et al. 2024; Maini et al. 2024) that show efficacy of "textbook-like" and "educational" data in pretraining LLMs, we use several knowledge-rich formats such as textbook entries, blog posts, wikihow, *inter alia* to enhance factual synthesis and structured reasoning. We also use several creative styles such as moral stories, poetry, reddit posts and others to encourage generative fluency and imagination. We list all prompts utilised in Appendix C.

3.2 Persona-Based Generation

We leverage PersonaHub (Ge et al. 2024), an open source repository of 371M personas, to synthetically generate 164.3M English Indic context personas. Additionally, we follow their approach to synthetically generate 50K Indic language personas, from Indic Web documents, that cover the diverse Indian linguistic, regional, and sociocultural identities. An example of a generated persona from our dataset: *A young software engineer from Bangalore who codes all day and hits the gym hard at night*.

We then use these personas as context in our synthetic generation pipeline following two approaches: (1) *Persona-based generation*, guided solely by the persona and language input to produce free-form, culturally fluent text; and (2) *Persona and document-based generation* where a persona is paired with a document, sampled either at random or based on its semantic similarity to the persona, for a more controlled and contextually rich generation.

Language	Generate in xx			Generate in En and translate to xx		
	gemma3	krutrim2	llama-3.3-70B	gemma3	krutrim2	llama-3.3-70B
Bengali	✓	✗	✗	✗	✗	✗
Gujarati	✗	✓	✗	✗	✗	✗
Hindi	✓	✓	✗	✗	✗	✗
Kannada	✗	✗	✗	✗	✓	✗
Malayalam	✗	✓	✓	✗	✗	✗
Marathi	✓	✓	✗	✗	✗	✗
Oriya	✗	✗	✓	✗	✗	✗
Punjabi	✓	✓	✓	✗	✗	✗
Tamil	✓	✓	✗	✓	✗	✗
Telugu	✓	✓	✗	✗	✗	✗

Table 1: Language-wise mapping of models used for direct generation (in language xx) and translation (generation in En followed by its translation to xx). Corresponding detailed human evaluations for the models are included in the Appendix in Table 11-21. Interestingly, LLaMA-3.3 70B (Grattafiori et al. 2024) showed superior performance than LLaMA-4 (Meta 2025) series for Indian languages.

3.3 Math and Reasoning-Based Synthetic Data

We introduce a novel methodology for generating high-quality pretraining data from existing instruction-tuning datasets for math and reasoning. Our method transforms existing, verified Question-Solution (Q-S) pairs from instruction-tuning datasets (Hendrycks et al. 2021b; Li et al. 2024a; Moshkov et al. 2025) into comprehensive and self-sufficient textbook sections. Specifically, we condition a generation model on a Q-S pair and instruct it to first introduce the underlying mathematical or technical concepts and theorems required to understand the problem, and then present a detailed, step-by-step solution. We posit that this approach offers two key advantages. Firstly, because the generation is grounded in an already-verified solution, it maintains the mathematical correctness and obviates the need for an additional, complex verification step. Secondly, we hypothesize that this "concept-then-solution" format will better equip models to emulate human-like reasoning

[2]https://github.com/huggingface/text-clustering

3.4 Topic-Aware Retrieval Augmented Generation (RAG)

To ensure extensive and accurate coverage of the Indian context, especially within long-tail topics, we first curate a detailed collection of Indic-specific topics. This is accomplished by systematically traversing the Wikipedia knowledge graph starting from the root node *Category:India*[3] up to a depth of three, resulting in a dataset containing over 10,000 topic titles. Next, we cluster our existing synthetic data using Vyakyarth[4]- a multilingual semantic embedding model tailored for Indic languages. We then filter the identified Indic topics by nearest neighbour based similarity score and subsequently applying a distance threshold. This ensures the retained topics are different from the topics already covered by the previously generated synthetic data. Finally, for each filtered topic, we utilize the SERP API[5] to retrieve relevant external documents. Leveraging these retrieved documents, we apply Retrieval Augmented Generation (RAG) techniques (Lewis, Perez et al. 2020) to generate contextually accurate and linguistically diverse content in multiple Indian languages.

3.5 Translation of English Synthetic Data

In addition to the different generation strategies discussed earlier, we also translate the 25B English synthetic *Cosmopedia* (Ben Allal et al. 2024) dataset, originally generated using the Mixtral-8x7B-Instruct-v0.1 model (Jiang, Sablayrolles et al. 2024). We evaluate various translation models across languages (Refer to Table 26 in the Appendix) and select Sarvam-Translate (Sarvam AI 2025b) for this translation. In order to ensure knowledge diversity across languages, we translate each of the 30M documents in *Cosmopedia* to only one randomly sampled Indic language.

4 Quality Evaluation Pipeline

Recent scaling laws (Chang et al. 2024; Chen et al. 2025) have argued the importance of quality data in pre-training. In order to assess the quality of synthetic data and filter out low-quality data at scale, we develop an automated quality evaluation pipeline comprising multiple heuristic and model-based filters outlined below.

1. Language consistency filter: Multilingual LLMs, especially when used in mid-to-low resource language settings, might generate text in mixed languages or in a language different from the intended language in the prompt. To ensure the generated data is in the target language, we leverage an ensemble language identification (LID) module optimized for Indian languages, building on top of the recent works (Khan et al. 2024).

2. Heuristic content filters: This module filters low-quality text in our generated large-scale corpora using rule-based heuristics and statistical features. It targets undesirable content such as NSFW material, excessive stopword or word repetition, anomalous characters (e.g., non-Latin/Indic scripts), outlier word counts, generic boilerplate, and references to third-party AI systems. Each criterion is governed by empirically tuned thresholds, and texts exceeding these limits are excluded to ensure high-quality data for downstream tasks, reported in Appendix (see Table 28).

3. Fluency evaluation (Perplexity filtering): In order to evaluate the fluency of the generated synthetic data, we train a 5-gram Kneser-Ney model (Heafield 2011).

The model is trained on 14.5M high-quality text samples sourced from Wikipedia, Sangraha (Khan et al. 2024), FineWeb2 (Penedo et al. 2024b), and bootstrapped synthetic corpora. For each data point, a perplexity score is computed and compared against language-specific thresholds, where low scores denote high linguistic coherence in the generated data. These thresholds are determined using held-out validation sets, with the 80th percentile of the score distribution used as the default cutoff, following earlier works (Khan et al. 2024). Further details regarding training and validation data used are in the Appendix (Table 29).

4. Quality classifiers: We also evaluate overall quality of the generated synthetic data on aspects such as content accuracy, clarity, coherence, grammatical correctness, informational depth, and overall usefulness, using a custom-trained FastText (Joulin et al. 2016) binary classifier to automatically assess the quality of Indic-language responses, labeling each instance as either `high` or `low` quality. The classifier is trained on approximately 384K examples labeled using the Gemini-1.5-Flash model through prompt-based evaluation. The training data comprises samples from diverse, high-quality sources such as Wikipedia, Sangraha (Khan et al. 2024), FineWeb2 (Penedo et al. 2024b), and generated synthetic corpora. The model achieves an overall accuracy of 98.9%, demonstrating high precision and recall across both quality classes on test split consisting of 160K examples. Details on training data composition, language-wise test set distribution, and evaluation metrics are provided in the Appendix (see Tables 30,31). On the source English documents, we also leverage pretrained models from NeMo Curator[6] library including the *FineWebEdu* classifier for detecting high-quality educational content and the *Domain Classifier* for categorizing the text into broad domains such as science, health, finance etc.

5. Bias detection and mitigation: We leverage the Word Embedding Association Test (WEAT) (Jentzsch et al. 2019) to quantify the social and cultural biases in our generated data. WEAT measures the strength of association between predefined *target* and *attribute* word sets in an embedding space, providing a quantitative estimate of implicit bias. The word embeddings are obtained from language-specific FastText models trained on our synthetic dataset. Our evaluation focuses on five key dimensions of social bias: gender, caste, race, religion, and regional/linguistic identity (Refer to Table 32 in the Appendix). For each dimension, we capture representative stereotypes using manually curated target and attribute word sets, each comprising 18–20 terms per language (Figures 5-9 show manually curated Hindi bias words across various bias aspects). Higher WEAT scores (typically

[3]https://en.wikipedia.org/wiki/Category:India

[4]https://huggingface.co/krutrim-ai-labs/Vyakyarth

[5]https://serpapi.com/

[6]https://github.com/NVIDIA/NeMo-Curator

Type	Generated Tokens (B)	Filtered Tokens (B)	Discard Rate (%)	Avg. source length	Avg. generated length
Document grounded - En (Section 3.1)	394.85	382.94	3.36	150	414
Document grounded - Indic (Section 3.1)	63.75	62.88	1.45	186	460
Persona based (Section 3.2)	37.83	37.24	1.56	34	242
Math & Reasoning (Section 3.3)	5.09	4.83	4.80	236	624
Topic based RAG (Section 3.4)	0.13	0.13	3.14	124	568
Translation (Section 3.5)	57.69	55.26	4.10	540	572

Table 2: Generated and filtered token counts (in billions) for each synthetic data source. Token counts are estimated using the LLaMA-4 tokenizer. We show discard rate as % of data filtered out and also report average output length (in words).

> 1.0) correspond to stronger stereotypical associations.

5 BhashaKritika: Synthetic Data

We used our pipeline to generate $\sim$ 540B tokens of high-quality synthetic data covering multiple Indian languages and Indic context topics. In Table 2, we show the distribution of this data by different sources used for generation. Here, "filtered tokens" correspond to the data that passes our quality evaluation pipeline and the "discard rate" is the percent of the synthetic data that is filtered out. Figure 2 illustrates the language-wise and topic-wise distribution of our synthetic data respectively. Each of these 12 topics in turn covers multiple Indic context sub-topics, for instance, *Indian culture and society* subsumes *Indian lifestyle*, *Indian philosophy*, *Indian fashion*, and others. We provide a comprehensive report of the different prompts, classifier datasets, annotation instructions as well as the quality evaluation in the Appendix for reproducibility.

6 Experiments

We conduct several ablations over the data sources, their language, the language of prompt instructions, and the personas to inform our choices in the synthetic generation process. Also, in order to evaluate the efficacy of our synthetic data in pretraining LLMs, we conduct experiments with a 1B parameter LLaMA-3.2 architecture in the compute constrained settings. We report the key findings here.

6.1 Does Language of Source Document and Prompt Impact the Quality of Generations?

Our synthetic data generation pipeline leverages source documents in both English (*e.g.* FineWeb) and Indian languages (*e.g.* FineWeb2) as the grounding context. *How does the language of this context impact the quality of generated synthetic data? For context in Indic languages, is it better to provide prompt instructions in English or Indic languages?* In order to answer these questions, we conduct an evaluation on the documents sampled from Pralekha (Suryanarayanan et al. 2024), a large-scale parallel document dataset in English and Indic languages, as context and the prompt instruction in English or Indic. We leverage our quality evaluation pipeline and report the discard rate on the generated synthetic data (Table 3). We observe that the models perform better when prompted with context in the same language as the intended language of generation and prompt instructions in English perform better than those in Indic. We, therefore, use prompts in English across all our synthetic generations.

Language	En/En	Ind/En	Ind/Ind
Bengali	0.57	0.47	**0.33**
Gujarati	17.45	**15.85**	24.15
Hindi	1.025	**0.6**	1.125
Malayalam	11.50	**9.70**	11.40
Marathi	1.20	**1.05**	0.93
Punjabi	7.97	5.03	**4.97**
Tamil	3.83	**3.00**	5.30
Telugu	3.10	**3.03**	4.73

Table 3: Impact of language of source document and prompt instructions on quality of generations (discard rates %). Ind/En denotes Indic document with English prompt.

Generation Mode	Discard rate%
Indic Persona	1.48%
En Persona	1.93%
En Persona with matching document	3.50%
En Persona with random Document	14.43%

Table 4: Impact of language of source persona and additional document grounding through a pilot study.

6.2 Does Language of Persona Impact the Quality of Generations?

We conduct a pilot study to evaluate the impact of additional document grounding by appending personas with either matching or random documents, selected using FAISS similarity scores. As shown in Table 4, this grounding significantly increases discard rates, suggesting that random pairing introduces linguistic inconsistencies and quality degradation. Additionally, generations conditioned on Indic personas yield lower discard rates compared to those grounded in English personas.

6.3 How Much Data is Filtered Out by the Quality Evaluation Pipeline and Why?

We ablate over the quality filters in the quality evaluation pipeline and present our insights. Refer to Table 27 in the Appendix for filter-wise discard rates across languages.

1. Language consistency filter: Language inconsistency is predominantly observed in Gujarati and Hindi languages (over 10% compared to an average of 7.6% across languages). This is primarily due to generations in other languages from the same language family (Marathi or Sanskrit instead of Hindi) and regional references in the document context. For instance, a news article mentioning Telangana Govt. used as context leads to generation in Telugu instead

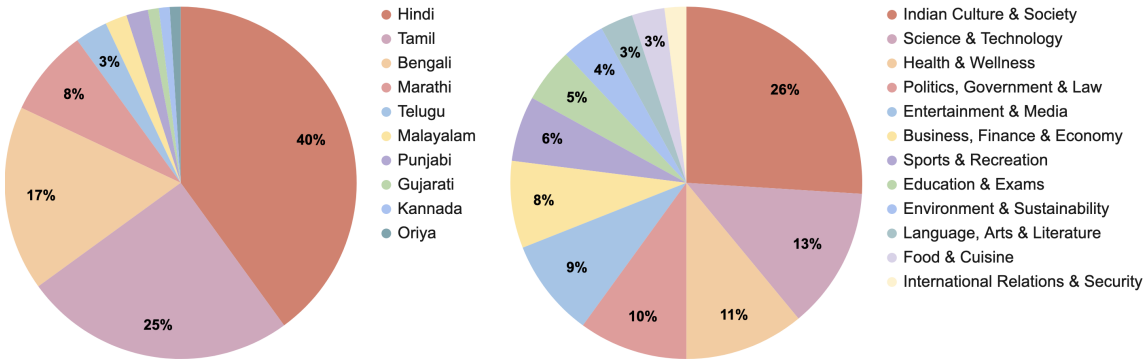

Figure 2: Distribution of languages and topics in BhashaKritika. We show the broad 12 topics for brevity with a more fine-grained distribution in Table 25 (Appendix).

of Gujarati (target language mentioned in the prompt).

2. Heuristic content filters: Length violations are the most common issue, affecting 2.26% generations, primarily due to incomplete or excessively verbose generations outside the 100–2500 word range. While toxic content is generally well-controlled, 1.13% of outputs contain NSFW material, and 2.02% include references to other AI systems. Word repetition (0.34%) and the use of excessive stopwords or non-Latin/non-Indic scripts (under 0.01%) remain rare. However, this filtering relies on manually curated keyword lists that are not exhaustive, and certain NSFW terms are context dependent, occasionally leading to false positives.

3. Fluency (perplexity-based) filter: Most generations are reasonably fluent, however, certain languages like Tamil and Bengali, show alarmingly high rates (above 10%). We observe the presence of English named entities and occasional English noun references impacting the perplexity scores, suggesting further room for improvement in our KenLM-based perplexity scoring.

4. Quality classifiers: Overall 3.40% of the total outputs are flagged as low quality by the quality classifiers. Some languages like Malayalam exhibit disproportionately high low-quality rates, primarily due to frequent word/character repetitions and poor linguistic coherence. While the classifier-based filter complements the content filters for repetition, NSFW, and stopwords, and the perplexity-based fluency filter, its key limitation is the dependency on domain-specific training data. Incorporating new styles or source types necessitates retraining the classifier unlike the relatively low-cost heuristic and statistical filters.

5. Bias detection: We evaluate our Hindi synthetic corpora across styles (*e.g.*, textbook, blogpost, persona) for Indian sociolinguistic bias. For each style, we report WEAT effect sizes and scores, computed over 1M samples, using curated target-attribute word sets. The analysis reveals consistent medium to high stereotypical bias across socio-cultural dimensions. Caste bias has effect sizes between 0.56–1.09, highest in persona. Gender bias is most pronounced in story (1.58) and Redditpost (1.21) styles, with high bias in four of seven styles. Race bias scored above 1.0 in most styles, peaking in blogpost (1.51), textbook (1.46), and Wikihow (1.28). Religion bias was similarly high in blogpost (1.39), textbook (1.3), and Wikihow (1.73) styles, indicating strong 'Hindu–Muslim' stereotypes. Region/linguistic bias was present but weaker, with translation showing a reverse effect, suggesting mitigation. These findings indicate prevalent and measurable biases in synthetic generations, especially regarding religion, race, and gender. The complete results are provided in the Appendix, Table 33.

6. Bias mitigation: We conduct a small-scale intervention targeting religious bias in Hindi textbook-style synthetic data. For around 20 biased instances per target term (*e.g.* Islamic and Hindu words), identified based on stereotypical co-occurrences, we replace them with LLM-based synthetically generated counter-stereotypical examples by reversing associations (*e.g.*, Islam association with positive and Hindu with negative attributes). Retraining FastText embeddings on this modified corpus reduced the WEAT effect size and score from (1.34, 1.11) to (1.29, 1.03). This finding suggests that this targeted data augmentation is a scalable mitigation strategy in our synthetic generation pipeline. Detailed results are available in the Appendix Figure 10.

7. Bias comparison (Web vs. BhashaKritika): We leverage documents from the Web as context in our synthetic generation pipeline. In order to evaluate the inherent bias mitigation in our pipeline, we compute WEAT scores on the source Web documents and the corresponding generated synthetic data (Refer to Table 34 in the Appendix). For instance, the religious bias in 'Hindi textbook-style' data, with effect size and WEAT score of (1.43, 1.35) in the source documents, dropped to (1.14, 0.93) in the generated synthetic data. These results indicate that our synthetic data has lower biases compared to those in the source Web data, with targeted interventions, as described in the last section, further aiding the debiasing. Detailed association scores for individual target words and other bias dimensions are provided in the Appendix (Figures 11-13).

6.4 How does Synthetic Data Compare to Web Data for LLM Training?

In addition to intrinsic data quality evaluation, we also evaluate our synthetic data for LLM pretraining. Starting from the pretrained checkpoint of LLaMA-3.2 1B model, we perform annealing (Grattafiori et al. 2024; Allal et al. 2025;

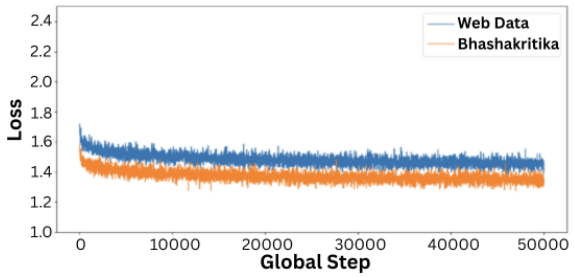

Figure 3: We annealed LLaMA-3.2 1B pretrained model on 50B tokens of Web vs. our synthetic data - BhashaKritika. We observe faster convergence on BhashaKritika.

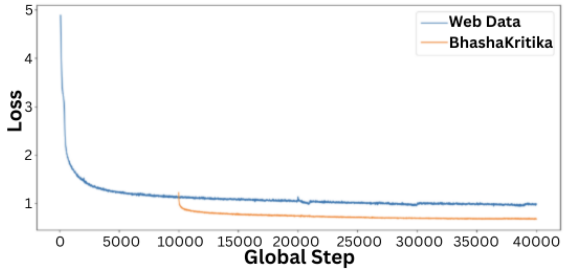

Figure 4: Loss curves for simulated low resource setting: LLaMA-3.2 1B is pretrained from scratch on 15B Indic Web tokens (10K training steps) followed by continual training on - (1) same Web data; (2) BhashaKritika data

OLMo et al. 2025), where we linearly decay LR to 0 over 50B tokens of training data comprising 70% Web, Math, and code data and 30% Indic data. We train two models - M_{Web} and M_{BK} where the Indic data is sampled from the Web and BhashaKritika, our Indic synthetic corpus, respectively. We attribute the faster convergence of M_{BK} (Fig. 3) to the high-quality and knowledge-dense nature of our synthetic data while the Web data tends to be relatively noisy (Abdin et al. 2024). In Table 5, we report the performance of these models on the English and Indic benchmarks. Further implementation details are provided in Appendix B.

Dataset	Web	BhashaKritika
Hellaswag	0.483	0.482
MMLU (Hendrycks et al. 2021a)	0.412	0.408
OpenbookQA	0.276	0.268
GSM8K (Cobbe et al. 2021)	0.111	**0.120**
DROP (F1) (Dua et al. 2019)	0.077	**0.083**
TriviaQA (Joshi et al. 2017)	0.398	**0.404**
ARC Easy	0.734	**0.741**
Winogrande (Sakaguchi et al. 2021)	0.616	**0.628**
ARC Challenge	0.406	**0.408**
CommonsenseQA (Talmor et al. 2019)	0.410	**0.411**
Indic Sentiment (Doddapaneni et al. 2023)	0.617	**0.631**
Indic Copa (Doddapaneni et al. 2023)	0.575	**0.588**
ARC Challenge Indic (Sarvam AI 2025a)	0.225	**0.225**
MILU (Verma, Khan et al. 2024)	0.283	0.282
Indic XNLI (Doddapaneni et al. 2023)	0.433	0.403
Indic XParaphrase (Doddapaneni et al. 2023)	0.782	0.736

Table 5: Evaluation results on English and Indic benchmarks for the LLaMA-3.2 1B pre-trained model annealed on 50B tokens of Web vs. BhashaKritika data. Results indicate that high-quality synthetic data can serve as an effective substitute for real-world data.

6.5 Can we Use Synthetic Data in Low Resource Settings?

A key challenge in building models for Indian languages is the limited availability of high-quality data. We explore whether our BhashaKritika corpus could serve as a good pretraining data in these low resource settings by conducting a controlled experiment. We pretrain LLaMA-3.2 1B model from scratch on a fixed budget of 15B tokens of Indic Web data (10K training steps in Fig. 4). Starting from this base model, we continually pretrain M_{Web} for 3 more epochs on the same Indic Web data and M_{BK} on data sampled from our BhashaKritika synthetic corpus.

The model trained on our Indic synthetic data converges faster and shows a similar or better performance on Indic benchmarks (Table 6). This indicates that high-quality synthetic data can serve as a viable substitute when Web data is limited, offering a promising direction for low-resource language settings.

Dataset	Web	BhashaKritika
Indic Sentiment (Doddapaneni et al. 2023)	0.491	**0.499**
MILU (Verma, Khan et al. 2024)	0.235	**0.236**
Indic Copa (Doddapaneni et al. 2023)	0.509	**0.512**
Indic XParaphrase (Doddapaneni et al. 2023)	0.499	**0.500**
Indic XNLI (Doddapaneni et al. 2023)	0.330	**0.339**
ARC Challenge Indic (Doddapaneni et al. 2023)	0.213	0.210

Table 6: Benchmark comparison on Indic datasets; 1B model pretrained on 15B tokens of Indic Web data from scratch is continually pretrained on multiple epochs of the same Web data vs BhashaKritika data.

7 Conclusion

We introduced *BhashaKritika*, a 540B tokens high-quality Indic synthetic corpus across 10 languages and different knowledge-dense styles. The data is generated using our scalable synthetic generation pipeline comprising multiple data sources, five generation approaches, multilingual LLMs, and translation models. We show that by careful selection of models per language and using Indic documents, topics and personas for grounding, we can synthetically generate high-quality Indic data. We demonstrate that using English instructions alongside Indic source texts yields better quality outputs and also introduce a novel technique to create math and reasoning focused data. Further, we introduce a comprehensive automated quality evaluation pipeline to ensure quality of the generated data. Through extensive analysis and empirical runs, we show the efficacy of our synthetically generated data, opening up avenues to augment the pretraining dataset for the low resource Indic languages.

Acknowledgements

We thank the leadership at Krutrim for their support in carrying out this research. We also thank the Data Annotation Team for their meticulous efforts in evaluation. We also thank the anonymous reviewers for their valuable feedback and suggestions.

References

Abdin, M.; Agarwal, S.; Awadallah, A.; et al. 2025. Phi-4-reasoning technical report. *CoRR abs/2504.21318*.

Abdin, M.; Aneja, J.; Behl, H.; Bubeck; et al. 2024. Phi-4 technical report. *CoRR abs/2412.08905*.

Agarwal, I.; Bozdag, N. B.; and Hakkani-Tür, D. 2025. Language Specific Knowledge: Do Models Know Better in X than in English? *CoRR abs/2505.14990*.

Allal, L. B.; Lozhkov, A.; Bakouch, E.; et al. 2025. SmolLM2: When Smol Goes Big–Data-Centric Training of a Small Language Model. *CoRR abs/2502.02737*.

Ben Allal, L.; Lozhkov, A.; Penedo, G.; Wolf, T.; and von Werra, L. 2024. Cosmopedia.

Chang, E.; Paltenghi, M.; Li, Y.; Lin; et al. 2024. Scaling Parameter-Constrained Language Models with Quality Data. *CoRR abs/2410.03083*.

Chen, Z.; Wang, S.; Xiao, T.; Wang; et al. 2025. Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In *ACL*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; et al. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Common Crawl. 2007. Common Crawl - Open Repository of Web Crawl Data.

Computer, T. 2023. RedPajama: an Open Dataset for Training Large Language Models.

Conneau, A.; et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *ACL*.

Doddapaneni, S.; Aralikatte, R.; Ramesh, G.; et al. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. arXiv:2212.05409.

Dua, D.; Wang, Y.; Dasigi, P.; et al. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. arXiv:1903.00161.

Gala, J.; et al. 2024. Airavata: Introducing Hindi Instruction-tuned LLM. *CoRR abs/2401.15006*.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR abs/2101.00027*.

Ge, T.; Chan, X.; Wang, X.; Yu, D.; Mi, H.; and Yu, D. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *CoRR abs/2406.20094*.

Grattafiori, A.; Dubey, A.; Jauhri; et al. 2024. The llama 3 herd of models. *CoRR abs/2407.21783*.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes; et al. 2023. Textbooks are all you need. *CoRR abs/2306.11644*.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proc workshop on statistical MT*.

Hendrycks, D.; Burns, C.; Basart, S.; et al. 2021a. Measuring Massive Multitask Language Understanding. In *ICLR*.

Hendrycks, D.; Burns, C.; Kadavath, S.; et al. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Jentzsch, S.; Schramowski, P.; Rothkopf, C.; and Kersting, K. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proc AAAI*.

Jiang, A. Q.; Sablayrolles, A.; et al. 2024. Mixtral of experts. *CoRR abs/2401.04088*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551.

Joulin, A.; Grave, E.; Bojanowski; et al. 2016. FastText.zip: Compressing text classification models. *CoRR abs/1612.03651*.

Kakwani, D.; Kunchukuttan, A.; Golla, S.; et al. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4948–4961. Online: Association for Computational Linguistics.

Kallappa, A.; Kamble, P.; Ravi, A.; et al. 2025. Krutrim LLM: Multilingual Foundational Model for over a Billion People. arXiv:2502.09642.

Khan, M. S. U. R.; Mehta, P.; Sankar; et al. 2024. IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages. *CoRR abs/2403.06350*.

Kohli, G. S.; Parida, S.; Sekhar, S.; et al. 2023. Building a Llama2-finetuned LLM for Odia Language Utilizing Domain Knowledge Instruction Set. arXiv:2312.12624.

Krutrim, T. 2024. Krutrim LLM: Multilingual Foundational Model for over a Billion People. *Under Review*.

Kunchukuttan, A.; Kakwani, D.; Golla, S.; et al. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *CoRR abs/2005.00085*.

Lambert, N.; Morrison, J.; Pyatkin, V.; et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. *CoRR abs/2411.15124*.

Lewis, P.; Perez, E.; et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proc NeurIPS*.

Li, J.; Beeching, E.; Tunstall, L.; et al. 2024a. NuminaMath.

Li, J.; Fang, A.; Smyrnis, G.; Ivgi, M.; Jordan, M.; et al. 2024b. DataComp-LM: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794.

Li, Y.; Bubeck, S.; et al. 2023. Textbooks are all you need ii: phi-1.5 technical report. *CoRR abs/2309.05463*.

Liu, R.; Wei, J.; Liu, F.; Si, C.; et al. 2024. Best practices and lessons learned on synthetic data for language models. *CoRR abs/2404.07503*.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*.

Maini, P.; Seto, S.; Bai, H.; et al. 2024. Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling. arXiv:2401.16380.

Meta, A. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Meyer, Y.; and Corneil, D. 2025. Nemotron-Personas: Synthetic Personas Aligned to Real-World Distributions.

Moshkov, I.; Hanley, D.; Sorokin, I.; et al. 2025. AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset. *CoRR abs/2504.16891*.

Muennighoff, N.; Rush, A.; Barak; et al. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36: 50358–50376.

Nadas, M.; Diosan, L.; and Tomescu, A. 2025. Synthetic data generation using large language models: Advances in text and code. *CoRR abs/2503.14023*.

Odumakinde, A.; D'souza, D.; Verga; et al. 2024. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress. *CoRR abs/2408.14960*.

OLMo, T.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; et al. 2025. OLMo 2 Furious. arXiv:2501.00656.

Penedo, G.; Kydlíček, H.; Lozhkov; et al. 2024a. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *CoRR abs/2406.17557*.

Penedo, G.; Kydlíček, H.; Sabolčec, V.; et al. 2024b. FineWeb2: A sparkling update with 1000s of languages.

Penedo, G.; Malartic, Q.; Hesslow, D.; et al. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *CoRR abs/2306.01116*.

Qin, Z.; Dong, Q.; Zhang, X.; Dong; et al. 2025. Scaling laws of synthetic data for language models. *CoRR abs/2503.19551*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*.

Sarvam. 2023. OpenHathi Series: An Approach To Build Bilingual LLMs Frugally.

Sarvam. 2024. Sarvam AI launches first LLM for Indian languages.

Sarvam AI. 2025a. Arc-Challenge-Indic.

Sarvam AI. 2025b. Sarvam-Translate.

Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; et al. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *ArXiv*, abs/2402.00159.

Su, D.; Kong, K.; Lin, Y.; Jennings, J.; et al. 2024. Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset. *CoRR abs/2412.02595*.

Suryanarayanan, S.; Song, H.; Khan; et al. 2024. Pralekha: An Indic Document Alignment Evaluation Benchmark. *CoRR abs/2411.19096*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. arXiv:1811.00937.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois; et al. 2023. Stanford alpaca: An instruction-following llama model.

Verma, S.; Khan, M. S. U. R.; et al. 2024. MILU: A Multi-task Indic Language Understanding Benchmark. *CoRR abs/2411.02538*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu; et al. 2022a. Self-instruct: Aligning language models with self-generated instructions. *CoRR abs/2212.10560*.

Wang, Y.; et al. 2022b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. CoRR abs/2212.10560.

Xu, C.; Sun, Q.; Zheng, K.; Geng; et al. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR abs/2304.12244*.

Xu, Z.; Jiang, F.; Niu, L.; et al. 2024. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. arXiv:2406.08464.

Yang, Y.; Patel, A.; Deitke, M.; Gupta; et al. 2025. Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation. *CoRR abs/2502.14846*.

Yu, Y.; Zhuang, Y.; Zhang, J.; Meng, Y.; et al. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *NeurIPS*.

Zhou, F.; Wang, Z.; Ranjan, N.; et al. 2025. MegaMath: Pushing the Limits of Open Math Corpora. arXiv:2504.02807.