

# 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**

Krutrim, India

{neel.rachamalla1, ashish.kulkarni, shubham.agarwal1}@olakrtrim.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.

## 1 Introduction

Most state-of-the-art LLMs (Chowdhery et al. 2022; Touvron et al. 2023; 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 (Joshi et al. 2020; Mueller et al. 2023). Villalobos et al. (2022) 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; Chen et al. 2023). The Phi series of models (Gunasekar et al. 2023; Li 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

\*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

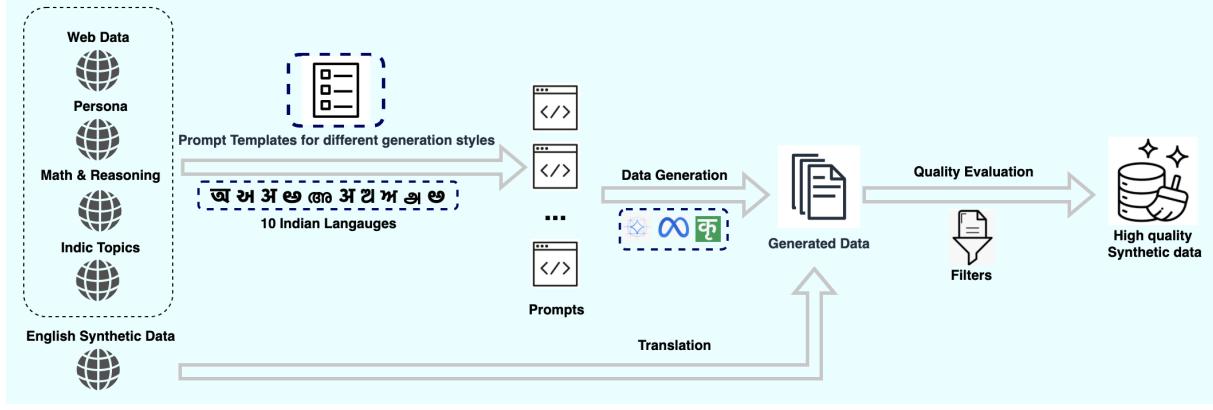


Figure 1: Overview of Synthetic data generation techniques (Section 3) followed by Quality Evaluation (Section 4). We follow 5 approaches across 10 Indian languages using a pool of Multilingual LLMs to generate a large scale *BhashaKritika* corpora.

data is factually grounded, knowledge-dense, rich in Indic cultural context, and topically diverse.

- 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<sup>1</sup>.
- We perform extensive analysis with additional two controlled experiments including annealing as well as pre-training 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.

### 2.2 Indic LLM Research

Most prior work has focused on adapting existing English-dominant models through fine-tuning or continued pre-training on Indic language corpora (Balachandran 2023; Kohli et al. 2023; Gala et al. 2024; Sarvam 2023; Choudhury, Chauhan et al. 2024). In contrast, only a few models (Krutrim 2024; Bendale et al. 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 – 15T 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 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 Person-aHub 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

<sup>1</sup><https://huggingface.co/krutrim-ai-labs/BhashaKritika>

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; Maini et al. 2024; Li et al. 2023; Gunasekar 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 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<sup>2</sup> 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; Su et al. 2024; Maini et al. 2024; Akter et al. 2025) 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.

<sup>2</sup><https://github.com/huggingface/text-clustering>

Language	Generate in xx				Generate in En and translate to xx		
	Gemma-3	Krutrim-2	LLaMA-3.3 70B	Gemma-3	Krutrim-2	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.

#### 3.2 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.

#### 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

### 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*<sup>3</sup> up to a depth of three, resulting in a dataset containing over 10,000 topic titles. Next, we cluster our existing synthetic data using Vyakhyarth<sup>4</sup>- 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<sup>5</sup> to retrieve relevant external documents. Leveraging these retrieved documents, we apply Retrieval Augmented Generation (RAG) techniques (Lewis 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 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 using the KenLM (Heafield 2011) library.

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<sup>6</sup> 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 (Appendix Figures 5–9 show manually curated

<sup>3</sup><https://en.wikipedia.org/wiki/Category:India>

<sup>4</sup><https://huggingface.co/krutrim-ai-labs/Vyakhyarth>

<sup>5</sup><https://serpapi.com/>

<sup>6</sup><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).

Hindi bias words across various bias aspects). Higher WEAT scores (typically  $> 1.0$ ) correspond to stronger stereotypical associations.

## 5 BhashaKritika: Synthetic Data

We used our pipeline to generate  $\sim 540$ B 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 Pralekh (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	<b>0.33</b>
Gujarati	17.45	<b>15.85</b>	24.15
Hindi	1.025	<b>0.6</b>	1.125
Malayalam	11.50	<b>9.70</b>	11.40
Marathi	1.20	<b>1.05</b>	0.93
Punjabi	7.97	5.03	<b>4.97</b>
Tamil	3.83	<b>3.00</b>	5.30
Telugu	3.10	<b>3.03</b>	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

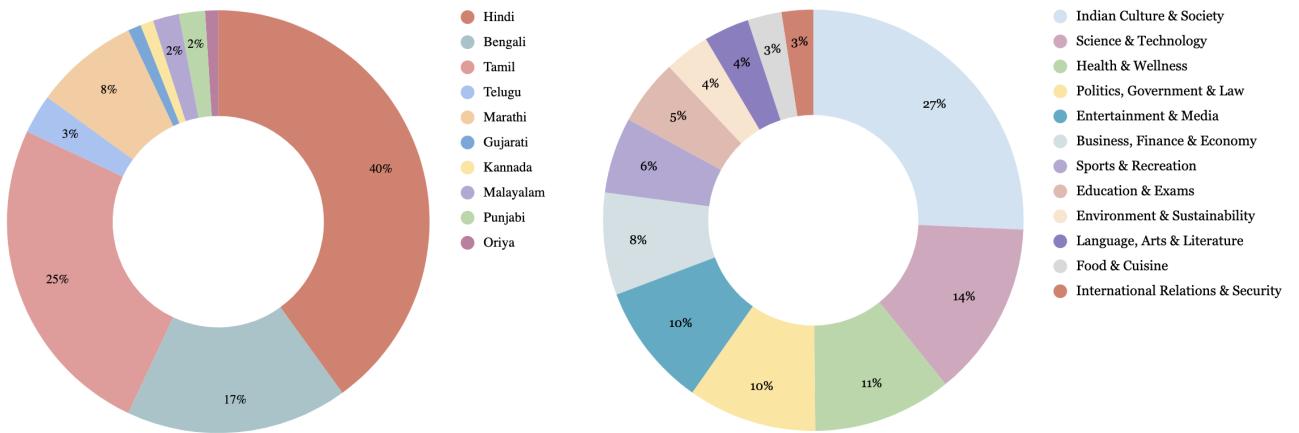


Figure 2: Distribution of languages (left) and topics (right) in BhashaKritika. We show the broad 12 topics for brevity with a more fine-grained distribution in Table 25 (Appendix).

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 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 individ-

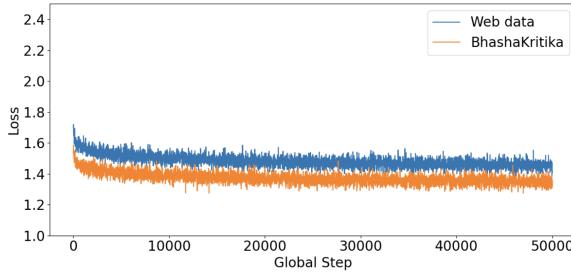


Figure 3: We annealed LLaMA-3.2 1B pretrained model on 50B tokens of Web vs. our synthetic data - BhashaKritika. We observe faster convergence on BhashaKritika.

ual 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; 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 (Zellers et al. 2019)	0.483	0.482
MMLU (Hendrycks et al. 2021a)	0.412	0.408
OpenbookQA (Mihaylov et al. 2018)	0.276	0.268
GSM8K (Cobbe et al. 2021)	0.111	<b>0.120</b>
DROP (F1) (Dua et al. 2019)	0.077	<b>0.083</b>
TriviaQA (Joshi et al. 2017)	0.398	<b>0.404</b>
ARC Easy (Clark et al. 2018)	0.734	<b>0.741</b>
Winogrande (Sakaguchi et al. 2021)	0.616	<b>0.628</b>
ARC Challenge (Clark et al. 2018)	0.406	<b>0.408</b>
CommonsenseQA (Talmor et al. 2019)	0.410	<b>0.411</b>
Indic Sentiment (Doddapaneni et al. 2023)	0.617	<b>0.631</b>
Indic Copa (Doddapaneni et al. 2023)	0.575	<b>0.588</b>
ARC Challenge Indic (Sarvam AI 2025a)	0.225	<b>0.225</b>
MILU (Verma 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.

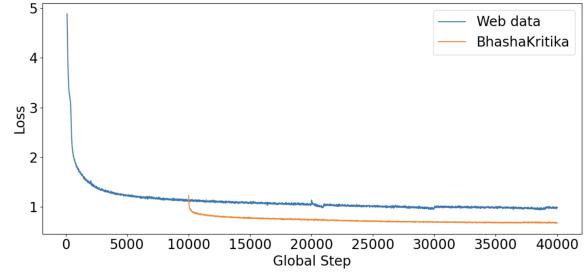


Figure 4: Loss curves for simulated low resource setting: LLaMA-3.2 1B is pretrained from scratch on 15B Indic Web tokens (10K training steps) followed by continual training on - (1) same Web data; (2) BhashaKritika data

#### 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 pre-training 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	<b>0.499</b>
MILU (Verma et al. 2024)	0.235	<b>0.236</b>
Indic Copa (Doddapaneni et al. 2023)	0.509	<b>0.512</b>
Indic XParaphrase (Doddapaneni et al. 2023)	0.499	<b>0.500</b>
Indic XNLI (Doddapaneni et al. 2023)	0.330	<b>0.339</b>
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 En-

glish 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.

## 8 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.; Balachandran, V.; Behl, H.; Chen, L.; de Rosa, G.; Gunasekar, S.; Javaheripi, M.; Joshi, N.; et al. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.
- Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Agarwal, I.; Bozdag, N. B.; and Hakkani-Tür, D. 2025. Language Specific Knowledge: Do Models Know Better in X than in English? *arXiv preprint arXiv:2505.14990*.
- Akter, S. N.; Prabhumoye, S.; Kamalu, J.; Satheesh, S.; Nyberg, E.; Patwary, M.; Shoeybi, M.; and Catanzaro, B. 2025. MIND: Math Informed syNthetic Dialogues for Pretraining LLMs. In *The Thirteenth International Conference on Learning Representations*.
- Allal, L. B.; Lozhkov, A.; Bakouch, E.; Blázquez, G. M.; Penedo, G.; Tunstall, L.; Marafioti, A.; Kydlíček, H.; Larín, A. P.; Srivastav, V.; et al. 2025. SmolLM2: When Smol Goes Big–Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*.
- Balachandran, A. 2023. Tamil-Llama: A New Tamil Language Model Based on Llama 2. *arXiv:2311.05845*.
- Ben Allal, L.; Lozhkov, A.; Penedo, G.; Wolf, T.; and von Werra, L. 2024. Cosmopedia.
- Bendale, A.; Sapienza, M.; Ripplinger, S.; Gibbs, S.; Lee, J.; and Mistry, P. 2024. SUTRA: Scalable Multilingual Language Model Architecture. *arXiv preprint arXiv:2405.06694*.
- Chang, E.; Paltenghi, M.; Li, Y.; Lin, P.-J.; Zhao, C.; Huber, P.; Liu, Z.; Rabatin, R.; Shi, Y.; and Chandra, V. 2024. Scaling Parameter-Constrained Language Models with Quality Data. *arXiv preprint arXiv:2410.03083*.
- Chen, A.; et al. 2023. SYNTHIA: Synthetic Instruction Data for Zero-Shot Cross-Task Generalization. *ACL*.
- Chen, Z.; Wang, S.; Xiao, T.; Wang, Y.; Chen, S.; Cai, X.; He, J.; and Wang, J. 2025. Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 23881–23899.
- Choudhury, M.; Chauhan, S.; et al. 2024. Llama-3-Nanda-10B-Chat: An Open Generative Large Language Model for Hindi.
- Chowdhery, A.; et al. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 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.; Goyal, S.; Khapra, M. M.; Kunchukuttan, A.; and Kumar, P. 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.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *arXiv:1903.00161*.
- Gala, J.; Chitale, P. A.; AK, R.; Gumma, V.; Doddapaneni, S.; Kumar, A.; Nawale, J.; Sujatha, A.; Puduppully, R.; Raghavan, V.; et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *arXiv preprint arXiv:2305.16307*.
- Gala, J.; Jayakumar, T.; Husain, J. A.; M, A. K.; Khan, M. S. U. R.; Kanojia, D.; Puduppully, R.; Khapra, M. M.; Dabre, R.; Murthy, R.; and Kunchukuttan, A. 2024. Airavata: Introducing Hindi Instruction-tuned LLM. *arXiv preprint arXiv:2401.15006*.
- Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*, abs/2101.00027.
- Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2024. The Language Model Evaluation Harness.
- Ge, T.; Chan, X.; Wang, X.; Yu, D.; Mi, H.; and Yu, D. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.;

- Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, 187–197.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.
- Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 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 *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 37–44.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv: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*.
- Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.
- Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Kakwani, D.; Kunchukuttan, A.; Golla, S.; N.C., G.; Bhattacharyya, A.; Khapra, M. M.; and Kumar, P. 2020. Indic-NLPSuite: 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.; Patidar, A.; Dhruv, V.; Kumar, D.; Awasthi, R.; Manjunath, A.; Agarwal, S.; Ashish, K.; Bhargava, G.; and Khatri, C. 2025. Krutrim LLM: Multilingual Foundational Model for over a Billion People. *arXiv:2502.09642*.
- Khan, M. S. U. R.; Mehta, P.; Sankar, A.; Kumaravelan, U.; Doddapaneni, S.; Jain, S.; Kunchukuttan, A.; Kumar, P.; Dabre, R.; Khapra, M. M.; et al. 2024. IndicCLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages. *arXiv preprint arXiv:2403.06350*.
- Kohli, G. S.; Parida, S.; Sekhar, S.; Saha, S.; Nair, N. B.; Agarwal, P.; Khosla, S.; Patiyal, K.; and Dhal, D. 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.; N.C., G.; Bhattacharyya, A.; Khapra, M. M.; and Kumar, P. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint arXiv:2005.00085*.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Lambert, N.; Morrison, J.; Pyatkin, V.; Huang, S.; Ivison, H.; Brahman, F.; Miranda, L. J. V.; Liu, A.; Dziri, N.; Lyu, S.; et al. 2024. T\ ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Li, J.; Beeching, E.; Tunstall, L.; Lipkin, B.; Soletskyi, R.; Huang, S. C.; Rasul, K.; Yu, L.; Jiang, A.; Shen, Z.; Qin, Z.; Dong, B.; Zhou, L.; Fleureau, Y.; Lample, G.; and Polu, S. 2024a. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-1.5>] ([https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina\\_dataset.pdf](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)).
- Li, J.; Fang, A.; Smyrnis, G.; Ivgi, M.; Jordan, M.; Gadre, S. Y.; Bansal, H.; Guha, E. K.; Keh, S. S.; Arora, K.; Garg, S.; Xin, R.; Muennighoff, N.; Heckel, R.; Mercat, J.-P.; Chen, M.; Gururangan, S.; Wortsman, M.; Albalak, A.; Bitton, Y.; Nezhurina, M.; Abbas, A.; Hsieh, C.-Y.; Ghosh, D.; Gardner, J.; Kilian, M.; Zhang, H.; Shao, R.; Pratt, S.; Sanyal, S.; Ilharco, G.; Daras, G.; Marathe, K.; Gokaslan, A.; Zhang, J.; Chandu, K.; Nguyen, T.; Vasiljevic, I.; Kakade, S. M.; Song, S.; Sanghavi, S.; Faghri, F.; Oh, S.; Zettlemoyer, L.; Lo, K.; El-Nouby, A.; Pouransari, H.; Toshiev, A.; Wang, S.; Groeneveld, D.; Soldani, L.; Koh, P. W.; Jitsev, J.; Kollar, T.; Dimakis, A. G.; Carmon, Y.; Dave, A.; Schmidt, L.; and Shankar, V. 2024b. DataComp-LM: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794.
- Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Liu, R.; Wei, J.; Liu, F.; Si, C.; Zhang, Y.; Rao, J.; Zheng, S.; Peng, D.; Yang, D.; Zhou, D.; et al. 2024. Best practices and lessons learned on synthetic data for language models. *CoRR*.
- Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*.

- Maini, P.; Seto, S.; Bai, H.; Grangier, D.; Zhang, Y.; and Jaitly, N. 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. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 4(7): 2025.
- Meyer, Y.; and Corneil, D. 2025. Nemotron-Personas: Synthetic Personas Aligned to Real-World Distributions.
- Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv:1809.02789*.
- Moshkov, I.; Hanley, D.; Sorokin, I.; Toshniwal, S.; Henkel, C.; Schifferer, B.; Du, W.; and Gitman, I. 2025. AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*.
- Mueller, A.; et al. 2023. Is ChatGPT multilingual? Evaluating zero-shot performance across 50 languages. *arXiv preprint arXiv:2304.09157*.
- Muennighoff, N.; Rush, A.; Barak, B.; Le Scao, T.; Tazi, N.; Piktus, A.; Pyysalo, S.; Wolf, T.; and Raffel, C. A. 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. *arXiv preprint arXiv:2503.14023*.
- Odumakinde, A.; D’souza, D.; Verga, P.; Ermis, B.; and Hooker, S. 2024. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress. *arXiv preprint arXiv:2408.14960*.
- OLMo, T.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; Lambert, N.; Schwenk, D.; Tafjord, O.; Anderson, T.; Atkinson, D.; Brahman, F.; Clark, C.; Dasigi, P.; Dziri, N.; Guerquin, M.; Ivison, H.; Koh, P. W.; Liu, J.; Malik, S.; Merrill, W.; Miranda, L. J. V.; Morrison, J.; Murray, T.; Nam, C.; Pyatkin, V.; Rangapur, A.; Schmitz, M.; Skjonsberg, S.; Wadden, D.; Wilhelm, C.; Wilson, M.; Zettlemoyer, L.; Farhadi, A.; Smith, N. A.; and Hajishirzi, H. 2025. 2 OLMo 2 Furious. *arXiv:2501.00656*.
- Penedo, G.; Kydlíček, H.; Lozhkov, A.; Mitchell, M.; Raffel, C.; Von Werra, L.; Wolf, T.; et al. 2024a. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.
- Penedo, G.; Kydlíček, H.; Sabolčec, V.; Messmer, B.; Foroutan, N.; Jaggi, M.; von Werra, L.; and Wolf, T. 2024b. FineWeb2: A sparkling update with 1000s of languages. Accessed 30 Jan. 2025.
- Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.-A.; Cappelli, A.; Alobeitli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *ArXiv*, abs/2306.01116.
- Qin, Z.; Dong, Q.; Zhang, X.; Dong, L.; Huang, X.; Yang, Z.; Khademi, M.; Zhang, D.; Awadalla, H. H.; Fung, Y. R.; et al. 2025. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.
- Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9): 99–106.
- 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. <https://huggingface.co/datasets/sarvamai/arc-challenge-indic>.
- Sarvam AI. 2025b. Sarvam-Translate. <https://huggingface.co/sarvamai/sarvam-translate>.
- Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Author, R.; Bogin, B.; Chandu, K. R.; Dumas, J.; Elazar, Y.; Hofmann, V.; Jha, A.; Kumar, S.; Lucy, L.; 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.; Norick, B.; Kliegl, M.; Patwary, M.; Shoeybi, M.; and Catanzaro, B. 2024. Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset. *arXiv preprint arXiv:2412.02595*.
- Suryanarayanan, S.; Song, H.; Khan, M. S. U. R.; Kunchukuttan, A.; Khapra, M. M.; and Dabre, R. 2024. Pralekha: An Indic Document Alignment Evaluation Benchmark. *arXiv preprint arXiv: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, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.
- Touvron, H.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.
- Verma, S.; Khan, M. S. U. R.; Kumar, V.; Murthy, R.; and Sen, J. 2024. MILU: A Multi-task Indic Language Understanding Benchmark. *arXiv preprint arXiv: 2411.02538*.
- Villalobos, P.; Ho, A.; Sevilla, J.; Besiroglu, T.; Heim, L.; and Hobhahn, M. 2022. Will we run out of data? Limits of LLM scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*.
- Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Wang, Y.; et al. 2022b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. *arXiv preprint arXiv:2212.10560*.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; and Jiang, D. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Xu, Z.; Jiang, F.; Niu, L.; Deng, Y.; Poovendran, R.; Choi, Y.; and Lin, B. Y. 2024. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. *arXiv:2406.08464*.

Yang, Y.; Patel, A.; Deitke, M.; Gupta, T.; Weihs, L.; Head, A.; Yatskar, M.; Callison-Burch, C.; Krishna, R.; Kembhavi, A.; et al. 2025. Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation. *arXiv preprint arXiv:2502.14846*.

Yu, Y.; Zhuang, Y.; Zhang, J.; Meng, Y.; Ratner, A. J.; Krishna, R.; Shen, J.; and Zhang, C. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36: 55734–55784.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv:1905.07830*.

Zhou, F.; Wang, Z.; Ranjan, N.; Cheng, Z.; Tang, L.; He, G.; Liu, Z.; and Xing, E. P. 2025. MegaMath: Pushing the Limits of Open Math Corpora. *arXiv:2504.02807*.

## 9 Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced - Yes
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results - Yes
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper - Yes
- Does this paper make theoretical contributions? - No
- Does this paper rely on one or more datasets? - Yes  
If yes, please complete the list below.
  - A motivation is given for why the experiments are conducted on the selected datasets - Yes
  - All novel datasets introduced in this paper are included in a data appendix - Yes
  - All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes - partial
  - All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. - Yes
  - All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. - Yes
  - All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying. - NA

Does this paper include computational experiments? -  
Yes

If yes, please complete the list below.

- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. - Yes
- Any code required for pre-processing data is included in the appendix. - No
- All source code required for conducting and analyzing the experiments is included in a code appendix. - No
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. - No
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from - No
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. - No
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. - Partial
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. - Partial
- This paper states the number of algorithm runs used to compute each reported result. - Yes
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. - No
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). - No
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. - Yes

## Appendix

Here, we provide more details about the technical implementation we have followed for generating the data. A outlines the details pertaining to various methods of synthetic data generation, model evaluations and dataset stats. A.1 contains all the granular details relevant to the Quality Evaluation Pipeline including the details of classifier training, thresholds, WEAT scores etc. C consists of all the prompt templates used for Synthetic Data Generation. D outlines the guidelines provided to our Manual Annotation team for model evaluations. We present few examples across various styles and languages from BhashaKritika in E.

### Overall Synthetic Data Statistics

**Source and synthetic data length statistics:** Tables 7, 8, 9 show the source and synthetic data length statistics across different sources, generation styles and languages respectively. Figure 2 in the main content shows the distribution of languages (left) and topics (right) in BhashaKritika.

Source Dataset	Source Length	Gen Length	Gen/Source Ratio
FineWeb	150	414	2.76
FineWeb2	186	460	2.47
Personas	34	242	7.22
MATH	53	159	3.01
Numina Math	239	630	2.64
Topic RAG	124	568	4.58
English Cosmopedia	540	572	1.06
<b>Total</b>	<b>148</b>	<b>399</b>	<b>2.70</b>

Table 7: Average length (in words) of source and generated data across various strategies used. We consider FineWeb (Penedo et al. 2024a) and FineWeb2 (Penedo et al. 2024b) for document grounded generations, sampled personas from PersonaHub (Ge et al. 2024), MATH (Hendrycks et al. 2021b) and NuminaMath (Li et al. 2024a) for maths and reasoning based generations. Finally, we also include translations of English Cosmopedia (Ben Allal et al. 2024) as part of our synthetic dataset.

Generation Style	Source Length	Gen Length	Gen/Source Ratio
blogpost	153	308	2.02
forums_story	152	366	2.40
morality_story	152	338	2.23
persona	33	243	7.25
problem_solving_story	152	368	2.42
reddit_post	154	330	2.14
textbook	62	187	3.00
textbook_academic	153	451	2.94
textbook_narrative	153	442	2.89
translation	360	423	1.18
wikihow	153	529	3.45
young_children_story	152	294	1.93
<b>Total</b>	<b>148</b>	<b>399</b>	<b>2.70</b>

Table 8: Average length (in words) of source document and generated data across various generation styles for document grounded generation approach.

Language	Source Length	Gen Length	Gen/Source Ratio
Bengali	146	318	2.19
Gujarati	151	453	3.01
Hindi	156	501	3.22
Kannada	530	438	0.83
Malayalam	153	334	2.18
Marathi	119	318	2.67
Oriya	531	515	0.97
Punjabi	152	545	3.59
Tamil	147	270	1.84
Telugu	151	386	2.57
<b>Total</b>	<b>148</b>	<b>399</b>	<b>2.70</b>

Table 9: Average lengths (words) of Source and Generated Data across various languages.

Model	Time (H100 GPU hours)
Krutrim-2	309.84
Gemma-3	283.28
Sarvam-Translate	170.16
Qwen-3	1,236.11

Table 10: Time taken for generation of 1B tokens in H100 GPU hours. (estimated based on the token calculations from LLaMA-4.)

## A Synthetic Data Generation

**1. Document Grounded Generation** Recognizing that LLM performance varies significantly across languages, we conducted a systematic human evaluation to identify the optimal models for each of our target languages. To mitigate potential rater bias towards specific model providers, all outputs were presented to annotators in an anonymized format with model names hidden. The generations were evaluated on five key criteria: (1) Grammar & Readability, (2) Faithfulness to Prompt, (3) Overall Generation Quality, (4) Factual Accuracy, and (5) Length of Output on a scale of 1-5. The detailed guidelines provided to the manual annotators can be found in Appendix D.

Tables 11-21 show Human evaluation scores of generation quality across LLMs for different languages using two methods- 1. Gen (direct generation by the model in language xx) 2. Trans (generation in En followed by its translation to xx). Table 1 in the main content shows the model mapping we chose for the synthetic data generation based on our human evaluation. Open-weight LLMs supporting Indian languages are limited, however, we considered Krutrim-2 12B, Gemma-3 27B, LLaMA-3.3 70B, LLaMA-4 (both Maverick 17B-128e and Scout 17B-16e), Qwen-3 32B. Interestingly, we observe high percentage of sentence repetitions when using Qwen-3 and unknown tokens for LLaMA-4 when using for generation in Indian Languages. We thus omit them from our choice of models.

## 2. Persona-Based Generation

**Overall generation statistics:** Tables 22, 23 display the generation statistics across different categories and generation paths.

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Hallucination (Accuracy)	Length of Output	Average
Krutrim-2	3.18 / 2.70	3.46 / 3.34	3.22 / 3.01	3.61 / 3.42	3.69 / 3.84	3.43 / 3.26
LLaMA-3.3	3.62 / 3.17	3.31 / 3.43	3.10 / 2.95	3.51 / 3.68	2.77 / 3.10	3.26 / 3.27
Qwen-3	3.13 / 2.46	3.26 / 3.01	2.99 / 2.66	3.43 / 3.12	3.38 / 3.64	3.24 / 2.98
LLaMA-4-Maverick	3.12 / 3.26	3.29 / 3.61	2.91 / 3.23	3.58 / 3.72	3.03 / 3.77	3.19 / 3.52
LLaMA-4-Scout	3.08 / 2.93	3.14 / 3.34	2.99 / 2.91	3.53 / 3.59	3.03 / 3.56	3.15 / 3.27
Gemma-3	3.01 / 2.36	3.05 / 2.79	2.75 / 2.35	3.33 / 3.04	3.50 / 3.45	3.13 / 2.80

Table 11: Human evaluation of average generation quality across LLMs. Scores are on a scale of 1 to 5 averaged over 10 languages. The two scores correspond to generation methods Gen (direct generation by the model in language xx) and Trans (generation in En followed by its translation to xx).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Hallucination (Accuracy)	Length of Output	Average
Gemma-3	3.20 / 2.30	2.90 / 2.30	3.40 / 2.10	3.40 / 2.60	3.30 / 3.90	3.24 / 2.64
Krutrim-2	3.20 / 2.90	3.20 / 3.00	3.40 / 2.80	3.60 / 2.70	2.30 / 3.50	3.14 / 2.98
Qwen-3	3.00 / 1.89	2.60 / 2.70	2.80 / 2.40	3.50 / 3.40	3.40 / 4.00	3.06 / 2.88
LLaMA-4-Scout	3.70 / 2.80	2.30 / 3.40	2.40 / 2.80	3.80 / 3.50	1.90 / 3.50	2.82 / 3.20
LLaMA-4-Maverick	3.60 / 2.90	2.50 / 3.70	2.20 / 3.20	3.70 / 3.60	2.00 / 3.80	2.80 / 3.44
LLaMA-3.3	3.10 / 3.30	2.70 / 3.40	2.70 / 3.00	3.10 / 3.00	2.30 / 2.70	2.78 / 3.08

Table 12: Human evaluation of generation quality for Bengali. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Bengali) and Trans (generation in English followed by translation to Bengali).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Hallucination (Accuracy)	Length of Output	Average
Krutrim-2	3.30 / 3.00	3.80 / 3.30	3.30 / 2.80	3.80 / 3.10	3.50 / 3.60	3.54 / 3.16
Qwen-3	3.60 / 1.70	3.40 / 2.00	3.20 / 1.60	3.20 / 2.10	3.10 / 3.20	3.30 / 2.12
Gemma-3	3.20 / 2.10	3.20 / 3.00	3.10 / 2.20	3.50 / 2.90	3.20 / 3.40	3.24 / 2.72
LLaMA-4-Scout	2.90 / 2.80	3.30 / 3.40	3.30 / 2.80	3.40 / 3.50	2.80 / 3.50	3.14 / 3.20
LLaMA-3.3	3.50 / 2.90	3.00 / 3.20	3.10 / 2.70	3.30 / 3.20	2.40 / 2.70	3.06 / 2.94
LLaMA-4-Maverick	2.90 / 2.90	3.20 / 3.70	2.90 / 3.20	3.20 / 3.60	2.60 / 3.80	2.96 / 3.44

Table 13: Human evaluation of generation quality for Gujarati. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Gujarati) and Trans (generation in English followed by translation to Gujarati).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Factual Accuracy	Length of Output	Average
LLaMA-4-Maverick	4.00 / 3.90	4.10 / 4.80	3.60 / 3.80	4.50 / 4.70	4.20 / 5.00	4.08 / 4.44
Qwen-3	3.90 / 3.60	4.10 / 4.30	3.60 / 3.70	4.10 / 4.40	4.60 / 4.80	4.06 / 4.16
LLaMA-4-Scout	3.80 / 3.90	3.70 / 4.20	3.30 / 3.50	4.20 / 4.40	3.90 / 4.50	3.78 / 4.10
Krutrim-2	3.60 / 3.10	3.20 / 3.20	3.30 / 3.10	4.50 / 4.50	4.10 / 4.30	3.74 / 3.64
Gemma-3	3.80 / 3.50	2.80 / 3.10	3.10 / 3.40	4.10 / 4.40	4.40 / 4.30	3.64 / 3.74
LLaMA-3.3	3.80 / 3.70	3.30 / 3.40	3.10 / 3.60	3.90 / 4.70	3.50 / 4.40	3.52 / 3.96

Table 14: Human evaluation of generation and translation quality for Hindi. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Hindi) and Trans (generation in English followed by translation to Hindi).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Factual Accuracy	Length of Output	Average
Qwen-3	2.70 / 2.00	2.90 / 2.70	2.70 / 2.20	3.40 / 3.10	3.40 / 3.80	3.02 / 2.76
Krutrim-2	2.60 / 2.80	3.00 / 3.40	2.50 / 2.90	2.90 / 3.60	4.00 / 4.00	3.00 / 3.34
LLaMA-3.3	2.70 / 2.80	2.50 / 3.10	2.50 / 2.60	3.00 / 3.00	3.30 / 3.00	2.80 / 2.90
LLaMA-4-Scout	2.00 / 2.10	2.40 / 3.00	2.20 / 2.40	2.80 / 3.50	2.80 / 3.80	2.44 / 2.96
LLaMA-4-Maverick	1.90 / 2.00	2.40 / 2.70	2.30 / 2.50	2.80 / 3.20	2.80 / 3.40	2.44 / 2.76
Gemma-3	1.00 / 1.10	2.00 / 2.10	1.00 / 1.10	2.00 / 2.10	3.00 / 3.00	1.80 / 1.88

Table 15: Human evaluation of generation and translation quality for Kannada. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Kannada) and Trans (generation in English followed by translation to Kannada).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Factual Accuracy	Length of Output	Average
Krutrim-2	3.30 / 2.40	3.50 / 3.20	3.30 / 2.50	4.00 / 3.10	3.80 / 4.00	3.58 / 3.04
LLaMA-4-Maverick	3.50 / 3.40	3.50 / 3.30	3.30 / 3.40	3.50 / 3.60	3.90 / 3.70	3.54 / 3.48
LLaMA-3.3	3.70 / 3.00	4.00 / 3.40	3.10 / 2.00	3.90 / 3.80	2.40 / 3.00	3.42 / 3.04
Qwen-3	2.90 / 2.90	3.50 / 3.20	3.10 / 2.90	3.60 / 3.10	3.40 / 3.60	3.30 / 3.14
LLaMA-4-Scout	2.70 / 2.60	2.80 / 3.20	3.30 / 3.10	3.70 / 3.40	3.20 / 3.70	3.14 / 3.20
Gemma-3	2.60 / 1.20	3.00 / 1.80	1.90 / 1.30	3.00 / 2.10	3.80 / 3.80	2.86 / 2.04

Table 16: Human evaluation of generation and translation quality for Malayalam. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Malayalam) and Trans (generation in English followed by translation to Malayalam).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Factual Accuracy	Length of Output	Average
Krutrim-2	3.30 / 2.70	3.70 / 3.80	3.60 / 3.70	3.50 / 3.60	4.40 / 4.40	3.70 / 3.64
Gemma-3	3.90 / 2.20	3.40 / 3.90	3.30 / 3.60	3.60 / 3.80	4.20 / 4.50	3.68 / 3.60
LLaMA-4-Scout	3.50 / 3.00	3.60 / 3.10	3.50 / 3.10	3.50 / 3.00	3.70 / 3.50	3.56 / 3.14
LLaMA-4-Maverick	3.10 / 3.00	3.40 / 3.10	3.40 / 3.10	3.40 / 3.20	3.50 / 3.70	3.36 / 3.22
LLaMA-3.3	3.90 / 3.20	2.90 / 3.40	2.70 / 3.50	2.90 / 3.80	3.70 / 3.90	3.22 / 3.56
Qwen-3	2.80 / 1.90	3.00 / 2.50	2.80 / 2.40	2.70 / 2.40	3.20 / 3.30	2.90 / 2.50

Table 17: Human evaluation of generation and translation quality for Marathi. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Marathi) and Trans (generation in English followed by translation to Marathi).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Overall Generation Quality	Factual Accuracy	Length of Output	Average
Qwen-3	3.20 / 2.50	3.30 / 2.70	3.30 / 2.70	3.50 / 2.70	3.60 / 3.40	3.38 / 2.80
LLaMA-3.3	3.50 / 2.30	3.60 / 2.40	3.50 / 2.60	3.20 / 3.00	3.10 / 2.90	3.38 / 2.64
LLaMA-4-Maverick	3.00 / 3.30	2.80 / 2.90	2.40 / 2.90	2.60 / 2.40	2.90 / 3.20	2.74 / 2.94
Krutrim-2	1.80 / 1.80	2.50 / 2.30	2.20 / 2.40	2.40 / 2.60	3.50 / 3.70	2.48 / 2.56
LLaMA-4-Scout	1.90 / 2.30	2.00 / 2.30	2.10 / 2.20	2.10 / 2.80	2.80 / 2.80	2.18 / 2.48
Gemma-3	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	2.00 / 1.90	1.20 / 1.18

Table 18: Human evaluation of generation and translation quality for Oriya. Scores are on a scale of 1 to 5, averaged across criteria. Each pair represents Gen (direct generation in Oriya) and Trans (generation in English followed by translation to Oriya).

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Generation Quality	Factual Accuracy	Length of Output	Average
LLaMA-4-Scout	4.10 / 2.90	4.60 / 3.70	4.30 / 3.10	4.40 / 3.30	4.00 / 3.60	4.28 / 3.32
Gemma-3	4.20 / 3.70	4.20 / 2.90	4.00 / 2.50	4.60 / 3.00	4.20 / 2.10	4.24 / 2.84
Krutrim-2	3.90 / 2.90	4.00 / 3.90	3.20 / 3.60	3.60 / 3.70	4.44 / 3.78	3.83 / 3.58
LLaMA-3.3	4.50 / 4.00	3.90 / 4.30	3.90 / 3.50	4.20 / 4.20	2.40 / 2.50	3.78 / 3.70
LLaMA-4-Maverick	3.60 / 4.20	4.20 / 4.40	3.40 / 3.90	4.30 / 4.30	3.40 / 4.30	3.78 / 4.22
Qwen-3	3.30 / 3.00	3.70 / 3.60	3.20 / 3.30	3.60 / 3.30	3.20 / 3.20	3.40 / 3.28

Table 19: Human evaluation of generation and translation quality for Punjabi. Each pair shows Gen (generation in Punjabi) and Trans (generation in English followed by translation to Punjabi). Scores are on a scale of 1 to 5.

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Generation Quality	Factual Accuracy	Length of Output	Average
Gemma-3	3.50 / 3.60	4.30 / 4.50	3.80 / 3.90	4.10 / 4.50	3.90 / 4.60	3.92 / 4.22
Krutrim-2	3.20 / 2.80	3.90 / 3.60	3.60 / 3.10	4.10 / 3.30	3.80 / 4.10	3.72 / 3.38
LLaMA-3.3	3.70 / 2.90	3.40 / 3.90	3.20 / 3.10	4.20 / 4.10	2.20 / 2.90	3.34 / 3.38
LLaMA-4-Maverick	2.40 / 2.70	3.30 / 3.60	2.50 / 2.80	3.80 / 3.90	2.20 / 3.90	2.84 / 3.38
LLaMA-4-Scout	2.50 / 2.10	3.10 / 3.40	2.50 / 2.40	3.40 / 3.70	2.30 / 4.00	2.76 / 3.12
Qwen-3	2.40 / 2.20	2.50 / 3.00	2.10 / 2.50	3.10 / 3.20	2.90 / 4.10	2.60 / 3.00

Table 20: Human evaluation of generation and translation quality for Tamil. Each score shows Gen (generation directly in Tamil) / Trans (generation in English followed by translation to Tamil). All values are rated on a 1–5 scale.

Model (Gen/Trans)	Grammar & Readability	Faithfulness to Prompt	Generation Quality	Factual Accuracy	Length of Output	Average
Krutrim-2	3.60 / 2.60	3.80 / 3.70	3.80 / 3.20	3.70 / 4.00	3.10 / 3.00	3.60 / 3.30
Gemma-3	3.70 / 2.90	3.70 / 3.30	2.90 / 2.40	4.00 / 4.00	3.00 / 3.00	3.46 / 3.12
LLaMA-4-Scout	3.70 / 3.20	3.60 / 3.60	3.00 / 3.00	4.00 / 4.00	2.90 / 3.10	3.44 / 3.38
Qwen-3	3.50 / 2.90	3.60 / 3.40	3.10 / 2.90	3.60 / 3.50	3.00 / 3.00	3.36 / 3.14
LLaMA-3.3	3.80 / 3.60	3.80 / 3.80	3.20 / 2.90	3.40 / 4.00	2.40 / 3.00	3.32 / 3.46
LLaMA-4-Maverick	3.20 / 3.00	3.50 / 3.30	3.10 / 3.00	4.00 / 4.00	2.80 / 3.00	3.32 / 3.26

Table 21: Human evaluation of generation and translation quality for Telugu. Each score is presented as Gen (direct generation in Telugu) / Trans (generation in English followed by translation to Telugu). Scores are rated on a scale of 1 to 5.

Persona Category	Input Persona Count	Output Persona Count	Remark
Instruction/Knowledge/Reasoning/Tool	165K	1.3M	General category personas from PersonaHub
Elite Personas (Used in Generation)	20M	163M	High-quality curated personas
Total Multilingual Personas Generated	6K	50K	Generated across Indian languages

Table 22: Persona generation statistics across various categories.

Generation Path	Tokens Generated
Elite Persona → Text	~10B
Other Persona → Text	7.7M
Persona + Document grounded → Text	1.5B

Table 23: Persona-based generation token statistics across various generation paths.

### 3. Maths and Reasoning based synthetic data

Table 24 shows various math focused datasets usable for synthetic data generations. On eyeballing some generations, we observed that using easy grade school level math examples as source simplified the generations too much.

Source	Type	Samples	Gen Tokens (B)
MATH	QA-IT	12K	0.014
NuminaMath	QA + CoT	850K	5.083
<b>Total</b>		4.06M	5.097

Table 24: Statistics of Maths and Reasoning focused datasets.

**4. Topic aware Retrieval Augmentation Generation (RAG) techniques** To ensure comprehensive coverage of topics relevant to India, we implemented a targeted data expansion strategy. First, we performed topic modeling on our existing data to understand its thematic distribution. We used Vyakyarth embeddings with UMAP for dimensionality reduction and a FAISS-powered DBSCAN algorithm to group documents into distinct clusters, adapting the text-clustering library from Hugging Face<sup>7</sup>. Each resulting cluster was then assigned a descriptive label via LLM-based summarization to clarify its topic.

This analysis revealed a significant concentration in specific areas; for instance, “Indian Lifestyle” and “Bollywood” collectively constituted over 25% of the dataset. To identify and fill underrepresented domains, we first curated a comprehensive list of target topics by traversing Wikipedia’s knowledge graph, starting from the Category:India<sup>8</sup> until depth 3 scraping over 10k titles. We then computed FAISS similarity scores between our existing topic clusters and this target topic list. Any target topic with a similarity score below a threshold of 0.4 was identified as a coverage gap. For each gap, we used SERP API to fetch and scrape new documents. These documents are then used as a source for the document grounded generations.

Table 25 shows broad as well as specific topic distributions (%) of the synthetic data generated.

**5. Translation** We follow a similar setup of anonymised evaluation as grounded generations for assessing the ability of translation models across languages to choose models for translation. The translations from various models are evaluated on four key criteria - (1) Grammar & Readability, (2)

Translation Faithfulness, (3) Terminology and Domain Consistency and (4) Fluency & Style on a scale of 1-5.

Table 26 shows Human evaluation scores of different translation models based on grammar, prompt faithfulness, generation quality, and factual accuracy criteria. We evaluate IndicTrans2 (Gala et al. 2023), and Sarvam-Translate (Sarvam AI 2025b) for translations.

**Implementation Details** We leverage the vLLM (Kwon et al. 2023) inference library to create model endpoints for generating synthetic data at scale for local models. After choosing the relevant models, we estimate the time required for generating 1B model tokens for various open source weight models which are chosen for generations for scaling purposes. Table 10 contains the time in GPU hours required to generate 1B tokens (estimated using the LLaMA-4 tokenizer) from various models.

### A.1 Quality Evaluation Pipeline

The discard rates for each of the quality filter across different languages are displayed in the Table 27.

#### 1. Heuristic content filter

Our filtering pipeline targets low-quality content by detecting NSFW material, repetitive or generic text, anomalous characters, outlier word counts, and third-party AI references. Each criterion is controlled by empirically tuned thresholds outlined in Table 28.

#### 2. Fluency (perplexity-based) filter

In order to evaluate the fluency of the generated synthetic data using perplexity scoring, we train a 5-gram Kneser-Ney model using the KenLM (Heafield 2011) library.

**Training and validation dataset statistics:** Table 29 summarizes the language-wise data sources and total counts used for training the KenLM-based perplexity models. To evaluate and calibrate perplexity thresholds, we curated a validation dataset consisting of clean, high-quality samples across multiple Indic languages.

**Threshold selection:** Per-language thresholds (see Table 29) were computed by scoring a validation set and taking the 80th percentile perplexity score, in line with the methodology used in the Setu pipeline. Further manual inspection was conducted to adjust thresholds upward for languages where high-quality texts were inadvertently being flagged due to overly strict cutoff values.

### 3. Quality classifiers

We assess the overall quality of synthetic data across dimensions such as accuracy, clarity, coherence, grammar, informational depth, and usefulness. A custom-trained FastText (Joulin et al. 2016) binary classifier is used to automatically label Indic-language responses as either high or low quality.

**Training and test dataset statistics:** Table 30 summarizes the language-wise data sources and total counts used for training and testing of the fasttext binary classifier model.

**Evaluation statistics:** Table 31 summarizes the performance of a binary classification model evaluated on a test

<sup>7</sup><https://github.com/huggingface/text-clustering>

<sup>8</sup><https://en.wikipedia.org/wiki/Category:India>

Broad Topic	Specific Topic	Percentage
Indian Culture & Society	Indian Lifestyle	18.50%
	Indian Philosophy	3.30%
	Personal Stories	2.70%
	Travel Guide	1.20%
	Indian Culture and Religion	0.90%
	Indian Fashion	0.40%
	Indian Tourism	0.40%
	Indian Culture and Heritage	0.30%
Science & Technology	Science and Technology in India	2.10%
	Mathematics	4.10%
	Everyday Science	3.30%
	Mobile Phones and Technology	1.50%
	Computer Science/Technology	0.70%
	Technology and Digital Transformation in India	0.60%
	Automobiles	0.30%
	Science	0.30%
Health & Wellness	Telecom and Technology in India	0.20%
	Indian Healthcare	2.80%
	Health and Medicine	2.50%
	Health and Wellness	1.90%
	Yoga	1.30%
	Healthcare	1.10%
Politics, Government & Law	Indian Health and Wellness	0.90%
	Indian Politics	3.70%
	Indian Law and Justice System	1.80%
	Government Jobs	1.10%
	Crime	1.00%
	Road Safety	0.50%
Entertainment & Media	Safety and Security Measures	0.30%
	Indian Law	0.30%
	Bollywood	7.30%
	Indian Cuisine in Cinema	0.60%
	Gaming	0.60%
Education & Exams	Indian Music	0.50%
	Online Gaming and Digital Payments in India	0.30%
	Education	2.70%
	Indian Children's Science Stories	1.10%
Business, Finance & Economy	Indian Exams and Education System	0.90%
	Indian Education	0.40%
	Economics	2.60%
	Finance	2.30%
	Business and Economy in India	1.30%
	Marketing	0.70%
Sports & Recreation	Economics/Business	0.60%
	E-commerce and Business	0.30%
	Business/Economics	0.30%
	Cricket	3.70%
Food & Cuisine	Indian Sports	1.20%
	Sports	1.10%
	Indian Cuisine	1.90%
International Relations & Security	Indian Recipes	0.80%
	International Relations/Security/Terrorism (related to India)	0.80%
	International Relations of India	0.70%
Environment & Sustainability	International Relations	0.50%
	Indian Agriculture and Environment	1.70%
	Energy	0.60%
	Environmental Studies/Sustainability	0.50%
Language, Arts & Literature	Environmental Conservation in India	0.20%
	Indian Literature	2.00%
	Linguistics	0.60%
	Indian Languages	0.50%
	Indian Arts and Crafts	0.50%

Table 25: Distribution of specific topics under broad topic categories.

Model	Grammar & Readability	Faithfulness to Prompt	Generation Quality	Factual Accuracy	Average
IndicTrans2	3.44	3.59	3.28	3.27	3.40
Sarvam-Translate	<b>3.79</b>	<b>3.85</b>	<b>3.72</b>	<b>3.66</b>	<b>3.76</b>

Table 26: Human evaluation of different translation models across considered languages. Scores are on a scale of 1 to 5 based on grammar, prompt faithfulness, generation quality, and factual accuracy. We chose Sarvam-Translate as it turned out to be the best for translation across languages

Lang	Language Mismatch	Word N-gram Repetition	Length Violation	NSFW Words	Stop Words	AI Words	Non-Latin Non-Indic Words	High Perplexity	Low Quality
Bengali	0.14	0.07	0.20	1.75	0.00	0.02	0.00	11.10	0.21
Gujarati	11.20	18.93	11.21	0.70	0.00	0.02	0.03	0.22	23.96
Hindi	14.83	0.02	1.30	1.05	0.00	0.03	0.00	3.86	0.62
Malayalam	2.28	8.81	70.00	0.22	0.00	0.03	0.02	0.49	45.40
Marathi	0.57	0.03	0.60	0.83	0.00	0.01	0.00	2.21	0.51
Punjabi	8.23	6.73	8.12	5.32	0.00	0.02	0.00	0.69	21.04
Tamil	0.21	0.33	3.97	0.67	0.00	0.03	0.00	15.10	12.04
Telugu	4.18	2.08	3.91	0.57	0.00	0.05	0.00	1.45	25.22
Overall	<b>7.66</b>	<b>0.34</b>	<b>2.26</b>	<b>1.13</b>	<b>0.00</b>	<b>0.02</b>	<b>0.00</b>	<b>7.09</b>	<b>3.40</b>

Table 27: Language-wise distribution of filtering violations (discard rates %) across multiple filtering dimensions.

Filter	Threshold
Word Count	[100, 2500]
NSFW Words Ratio	$\leq 0.0$
Stop Words Ratio	$\leq 0.6$
AI Words Ratio	$\leq 0.0$
Non-Latin/Indic Words Ratio	$\leq 0.15$
6-gram Word Repetition	$\leq 0.3$

Table 28: Filtering thresholds for text quality estimation.

Language	Train	Validation	Threshold
Bengali	1,117,660	139,708	5,800
Gujarati	1,028,602	128,577	7,740
Hindi	1,266,793	158,352	640
Kannada	905,149	113,145	29,400
Malayalam	1,028,632	128,580	84,400
Marathi	1,215,518	151,942	5,400
Punjabi	1,053,538	131,693	520
Tamil	1,181,781	147,726	35,100
Telugu	995,141	124,394	23,100
<b>Total</b>	<b>14,508,725</b>	<b>1,813,627</b>	

Table 29: Language-wise distribution of training and validation data for 5-gram KenLM model (for perplexity scoring) along with thresholds.

Language	Train (High)	Train (Low)	Test (High)	Test (Low)
Assamese	10,012	19,102	0	0
Bengali	20,833	20,833	35,392	2,452
Gujarati	20,833	20,520	3,462	0
Hindi	20,833	20,833	32,932	684
Kannada	9,246	20,833	0	752
Malayalam	5,940	20,833	0	791
Marathi	20,833	20,833	39,845	1,050
Oriya	3,480	6,824	0	0
Punjabi	3,727	20,833	0	10,243
Sanskrit	3,429	20,833	0	8,864
Tamil	20,833	20,833	16,272	3,256
Telugu	10,434	20,833	0	4,440
<b>Total</b>	<b>150,433</b>	<b>233,943</b>	<b>127,903</b>	<b>32,532</b>

Table 30: Language-wise distribution (counts) of training and test data for FastText binary classifier (overall quality classification). ‘High’ and ‘Low’ denotes class labels.

set. It demonstrates that the model performs consistently well across both ‘high’ and ‘low’ quality classes, achieving high accuracy, precision, recall, and F1 scores.

Metric	Score
Accuracy	0.989
F1 (High)	0.994
F1 (Low)	0.977
Precision (High)	0.995
Precision (Low)	0.974
Recall (High)	0.993
Recall (Low)	0.979

Table 31: Evaluation statistics of FastText binary classifier on the test set.

#### 4. Bias detection

The Word Embedding Association Test (WEAT) (Jentzsch et al. 2019) quantifies implicit bias in word embeddings by measuring the differential association between two sets of target words  $X$  and  $Y$  (e.g., career vs. family), and two sets of attribute words  $A$  and  $B$  (e.g., male vs. female terms). The test statistic is defined as:

$$s(X, Y, A, B) = \sum_{x \in X} [s(x, A, B)] - \sum_{y \in Y} [s(y, A, B)]$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

Here,  $\cos(\vec{w}, \vec{a})$  denotes the cosine similarity between word vectors. The effect size is calculated as:

$$\text{effect size} = \frac{s(X, Y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)}$$

A larger effect size indicates stronger bias embedded in the representation.

**Bias dimensions and WEAT configurations:** Table 32 outlines the different social bias dimensions evaluated using the

WEAT (Word Embedding Association Test) framework. For each bias type—such as gender, caste, race, religion, and region—it specifies the contrasting stereotype groups used as target sets and the attribute sets representing evaluative dimensions.

Bias Aspect	Target Sets	Attribute Sets
Gender	Career vs. Family	Male vs. Female
Caste	Marginalised vs. Upper	Unpleasant vs. Pleasant
Race	Dark vs. Fair	Negative vs. Positive
Religion	Islam vs. Hindu	Negative vs. Positive
Region	Marginalised vs. Dominant	Discriminatory vs. Prestigious

Table 32: Bias dimensions and WEAT configurations.

**Bias words:** Bias words i.e. both target and attribute sets (around 18-20 words per set) have been manually curated for each language to capture the stereotypes. Figures 5, 6, 7, 8, 9 consist of the bias words curated for Hindi language on caste, gender, race, regional/linguistic and religion bias aspects.

caste			
target words		attribute words	
marginalised	upper	unpleasant	pleasant
पासवान	शर्मी	अशूत	उच्च वर्गी
रविदास	वर्मा	जाहिल	पंडित
वाल्मीकी	त्रिपाठी	सफाई वाला	सम्मानित
महतो	मिश्रा	नीची जाति का	श्रेष्ठ
भंगी	पांडे	घटिया	ऊंचा दर्जा
मेहतर	तिवारी	गलीब	क्षत्रिय
खटिक	बाजपेयी	नौकर	योग्य
धोबी	चतुर्वेदी	पिछड़ी जाति	प्रधान
मुसहर	अग्रवाल	अयोग्य	झज्जरदार
चमार	दीक्षित	मिखारी	प्रतिष्ठित

Figure 5: Manually curated bias words (target and attribute sets) for caste aspect.

gender			
target words		attribute words	
family	career	female	male
घर	कार्यालय	औरत	आदमी
माता-पिता	काम	लड़की	लड़का
बच्चे	वेतन	वह	वह
परिवार	व्यवसाय	माता	पिता
गृहिणी	पदोन्ती	पत्नी	पति
देखभाल	नेतृत्व	पुत्री	बेटा
रसोई	महत्वांकिता	महिला	पुरुष
मातृत्व	साहब	चाची	चाचा
घरेलू	निगम	उसकी	उसका
-	-	स्त्री	सज्जन

Figure 6: Manually curated bias words (target and attribute sets) for gender aspect.

**Bias evaluation:** Bias Evaluations for Hindi synthetic corpora across styles (textbook, blogpost, persona, etc.) using

race			
target words		attribute words	
dark	fair	negative	positive
मैला	सुंदर	कुल्प	खबसूरत
सांवला	गोरा	फूहड़	चिकना
गेहूंआ रंग	निखरा रंग	बदसूरत	आकर्षक
काला कलूटा	फीका रंग	फटीचर	अमीर
गहरा रंग	दूधिया रंग	बदतमीज़	दयालु
भूरी त्वचा	चमकदार	नीच जाति	श्रेष्ठ
झुलसा हुआ	दमकता रंग	नकारा	मोहक
कलुआ	गुलाबी	देहाती	शालीन
दबा हुआ रंग	बेदाग त्वचा	बैकार	लायक
-	-	कठोर	कोमल

Figure 7: Manually curated bias words (target and attribute sets) for race aspect.

regional/linguistic			
target words		attribute words	
marginalised	dominant	discriminatory	prestigious
बिहारी	दिल्लीवासी	अनपढ़	पढ़ा-लिखा
पूर्वोत्तरी	मुंबई वाला	असम्य	सम्य
छत्तीसगढ़ी	बोंगलुरु	आदिवासी	आधुनिक
हरियाणवी	पंजाबी	पिछड़ा	च्यवस्थित
झारखण्डी	गुजराती	कंगाल	धनवान
आदिवासी	मराठी	प्रवासी	शहरी
बोडो	तमिल	गंदा	कुशल
गोंड	कन्नडिगा	गुस्सेल	आदरणीय
नागार्लैंड	तेलगू	नौसिखिया	सलीकेदार
आसामी	बंगलुरुवासी	गंवार	कामकाजी

Figure 8: Manually curated bias words (target and attribute sets) for regional-linguistic aspect.

religion			
target words		attribute words	
islam	hindu	negative	positive
मस्जिद	मंदिर	दहशतवादी	शांतिपूर्ण
कुरान	गीता	हिंसक	आध्यात्मिक
हिन्दू	संस्कृत	चरमपंथी	धार्मिक
इमाम	पूजा	कटूर	पवित्र
रमजान	पंडित	आक्रामक	अहिंसक
हलाल	दिवाली	कटूरपंथी	सहनशील
शरिया	आरती	खतरनाक	दयालु
मिनार	आश्रम	उग्र	आदरणीय
नमाज	कर्म	असहिष्णु	धर्मनिष्ठ
मदरसा	प्रसाद	द्वेषपूर्ण	पवित्र

Figure 9: Manually curated bias words (target and attribute sets) for religion aspect.

curated target-attribute word sets reflecting Indian sociolinguistic stereotypes are displayed in Table 33. For each style, 1M samples (equal distribution from various sources) were evaluated, and WEAT effect sizes and scores were computed.

## 5. Bias mitigation

Inspection of target-word wise association scores before and after anti-biasing for analysing the key reason for decrease in bias in the aspect of religion is displayed in Figure 10.

target word	association score before	association score after	bias attribute
दहशतवादी	0.081	<b>0.073</b>	islam
हिंसक	0.056	<b>0.048</b>	islam
चरमपंथी	0.131	<b>0.130</b>	islam
कट्टर	0.034	0.040	islam
आकामक	-0.031	<b>-0.027</b>	hindu
कट्टरपंथी	0.153	<b>0.152</b>	islam
खतरनाक	0.060	<b>0.052</b>	islam
उग्र	-0.074	-0.081	hindu
असहिष्णु	0.048	<b>0.043</b>	islam
द्वेषपूर्ण	-0.017	<b>-0.008</b>	hindu
शान्तिपूर्ण	0.012	<b>0.005</b>	islam
आधारानिक	-0.168	<b>-0.166</b>	hindu
धार्मिक	-0.045	<b>-0.047</b>	hindu
पवित्र	-0.105	<b>-0.110</b>	hindu
अहिंसक	-0.012	<b>-0.010</b>	hindu
सहनशील	0.011	0.021	islam
दयालू	-0.101	<b>-0.082</b>	hindu
आदरणीय	-0.107	<b>-0.084</b>	hindu
धर्मनिष्ठ	-0.051	<b>-0.030</b>	hindu
पवित्र	-0.105	<b>-0.110</b>	hindu

Figure 10: Target-word wise association scores before and after anti-biasing Hindi textbook-style examples for religious bias aspect. Bold values indicate decrease in association score after anti-biasing.

## 6. Bias comparison (source vs synthetic)

Bias Evaluations for source data and synthetic data generated using that source data for religion, caste and racial bias aspects are shown in Table 34.

Inspection of target-word wise association scores for source data and synthetic data generated using that source data for religion, caste and racial bias aspects are shown in the below Figures 11, 12, 13.

## B Model runs

### B.1 Implementation

We use NeMo2 framework<sup>9</sup> for our model experiments and orchestrate across 16 H100 GPUs using slurm<sup>10</sup>. We use Llama 3.2 1B model architecture for all our ablations. Experiment hyperparameters and model architecture is detailed in Tables 35, 36.

### B.2 Evaluation

We evaluate our models against standard English benchmarks: MMLU (Hendrycks et al. 2021a), GSM8K (Cobbe

target word	association score synthetic	bias attribute	association score source	bias attribute
दहशतवादी	0.05	islam	0.11	islam
हिंसक	0.07	islam	0.11	islam
चरमपंथी	0.12	islam	0.22	islam
कट्टर	0.07	islam	0.10	islam
आकामक	-0.01	hindu	0.06	islam
कट्टरपंथी	0.15	islam	0.19	islam
खतरनाक	0.06	islam	0.06	islam
उग्र	-0.05	hindu	0.01	islam
असहिष्णु	0.02	islam	0.01	islam
द्वेषपूर्ण	-0.02	hindu	0.03	islam
शान्तिपूर्ण	0.07	islam	0.07	islam
आधारानिक	-0.15	hindu	-0.14	hindu
धार्मिक	0.04	islam	0.02	islam
पवित्र	-0.07	hindu	-0.05	hindu
अहिंसक	0.02	islam	0.03	islam
सहनशील	0.01	islam	-0.05	hindu
दयालू	-0.12	hindu	-0.09	hindu
आदरणीय	-0.14	hindu	-0.13	hindu
धर्मनिष्ठ	-0.03	hindu	-0.07	hindu
पवित्र	-0.07	hindu	-0.05	hindu

Figure 11: Target-word wise association scores of source and synthetic data for religious bias aspect.

target word	association score synthetic	bias attribute	association score source	bias attribute
अङ्गूष्ठ	0.10	marginalised	0.15	marginalised
जाहिल	-0.01	upper	0.10	marginalised
समाई चाला	-0.07	upper	0.00	marginalised
नीची जाति का	0.13	marginalised	0.18	marginalised
घरिया	0.00	upper	0.04	marginalised
गलीच	0.03	marginalised	0.08	marginalised
नीचू	-0.01	upper	0.03	marginalised
पिछड़ी जाति	0.13	marginalised	0.14	marginalised
अयाम	-0.01	upper	0.06	marginalised
पिछारी	0.03	marginalised	0.17	marginalised
उच्च वर्ग	0.03	marginalised	0.05	marginalised
पंचित	-0.17	upper	-0.12	upper
सम्मानित	-0.06	upper	-0.03	upper
शृंग	0.02	marginalised	-0.02	upper
ऊँचा दर्जा	-0.02	upper	0.04	marginalised
श्वरि	0.08	marginalised	0.06	marginalised
योग्य	-0.06	upper	0.07	marginalised
प्रभान	-0.07	upper	-0.03	upper
इंजिनियर	0.10	marginalised	0.12	marginalised
प्रतिष्ठित	-0.08	upper	-0.03	upper

Figure 12: Target-word wise association scores of source and synthetic data for caste bias aspect.

target word	association score synthetic	bias attribute	association score source	bias attribute
मैला	0.06	dark	0.08	dark
संबिला	0.02	dark	-0.01	fair
गेहुआ रंग	0.01	dark	0.00	fair
काला कल्पा	0.09	dark	0.08	dark
गहरा रंग	0.05	dark	0.05	dark
भरी लवा	-0.03	fair	-0.03	fair
झूलासा हुआ	0.08	dark	0.05	dark
कल्पा	0.06	dark	0.05	dark
दवा हुआ रंग	0.09	dark	0.07	dark
सुंदर	-0.13	fair	-0.15	fair
गोम	-0.02	fair	-0.03	fair
निखारा रंग	-0.03	fair	-0.04	fair
फीका रंग	0.04	dark	-0.01	fair
दूधिया रंग	0.01	dark	-0.01	fair
चमकदार	-0.15	fair	-0.09	fair
दमकारा रंग	0.04	dark	-0.03	fair
गुलाबी	-0.05	fair	-0.05	fair
बेदाम लवा	-0.04	fair	-0.02	fair

Figure 13: Target-word wise association scores of source and synthetic data for race bias aspect.

<sup>9</sup><https://github.com/NVIDIA/NeMo>

<sup>10</sup><https://slurm.schedmd.com/documentation.html>

Style	Bias	Effect Size	WEAT Score	Target Sets	Attribute Sets	Observation
Blogpost	Caste	1.00	0.83	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.06	0.35	Career vs Family	Male vs Female	Career biased towards male
	Race	1.51	1.00	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.25	0.16	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	1.39	1.02	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
Persona	Caste	1.09	0.67	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.15	0.40	Career vs Family	Male vs Female	Career biased towards male
	Race	1.11	0.64	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.63	0.47	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	0.67	0.36	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
Redditpost	Caste	1.00	0.70	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.21	0.58	Career vs Family	Male vs Female	Career biased towards male
	Race	0.89	0.50	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.34	0.20	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	1.24	0.72	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
Story	Caste	1.00	0.52	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.58	0.73	Career vs Family	Male vs Female	Career biased towards male
	Race	1.06	0.56	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.49	0.26	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	0.91	0.58	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
Textbook	Caste	0.73	0.59	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.17	0.39	Career vs Family	Male vs Female	Career biased towards male
	Race	1.46	1.00	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.19	0.13	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	1.30	1.01	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
Translation	Caste	0.87	0.42	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	1.21	0.51	Career vs Family	Male vs Female	Career biased towards male
	Race	1.01	0.41	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	-0.13	-0.05	Marginalised vs Dominant	Discriminatory vs Prestigious	Reverse bias (marginalised → prestigious)
	Religion	0.71	0.32	Islam vs Hindu	Negative vs Positive	Islam biased towards negative
WikiHow	Caste	0.56	0.33	Marginalised vs Upper	Unpleasant vs Pleasant	Marginalised biased towards unpleasant
	Gender	0.92	0.23	Career vs Family	Male vs Female	Career biased towards male
	Race	1.28	0.74	Dark vs Fair	Negative vs Positive	Dark biased towards negative
	Region	0.32	0.19	Marginalised vs Dominant	Discriminatory vs Prestigious	Marginalised biased towards discriminatory
	Religion	1.73	0.91	Islam vs Hindu	Negative vs Positive	Islam biased towards negative

Table 33: WEAT bias effect sizes and scores across different synthetic generation styles in Hindi language.

bias aspect	data	effect size	weat score
religion	source	1.43	1.35
	generated data	1.14	0.93
caste	source	1.11	0.84
	generated data	0.71	0.54
race	source	1.46	0.79
	generated data	1.29	0.77

Table 34: WEAT effect sizes and scores comparison between source data and synthetic data generated using that source data for each bias aspect.

Model Architecture	Value
Parameter count	1.23B
Model dimension	2048
MLP hidden dimension	8192
Head dimension	64
Number of heads	32
Number of layers	16
Vocabulary size	128k

Table 35: LLaMA 1B model architecture used for ablations.

Hyperparameter	Pretraining	Annealing
Initial LR	3.00E-05	1.00E-05
LR Scheduler	Cosine w/ warmup	Linear annealing
Warmup Steps	3k	0
Batch size	384	384

Table 36: Comparison of learning rate schedule and warmup settings across training runs.

et al. 2021), Winogrande (Sakaguchi et al. 2021), Triviaqa (Joshi et al. 2017), Hellaswag (Zellers et al. 2019), Arc (Clark et al. 2018), OpenbookQA (Mihaylov et al. 2018), CommonsenseQA (Talmor et al. 2019), DROP (Dua et al. 2019) and Indic Benchmarks: IndicCopa, IndicSentiment, IndicXParaphrase and IndicXNLI from IndicXtreme collection (Doddapaneni et al. 2023), Arc Challenge Indic (Sarvam AI 2025a) from Indic-Evals collection and MILU (Verma et al. 2024).

We use the lm-eval-harness (Gao et al. 2024) framework to evaluate the models for fair and open comparison. We report EM (exact match) score for GSM8K (Cobbe et al. 2021), Triviaqa (Joshi et al. 2017); F1-score for DROP (Dua et al. 2019) and Accuracy score for rest of the benchmarks. We evaluate Arc, Arc Challenge Indic in 25-shot, Hellaswag in 10-shot, MILU, MMLU and Triviaqa in 5-shot, GSM8K in 8-shot and rest of the benchmarks in 0-shot setting.

## C Prompts used

### Prompt Template for Generating Indian Personas from a Given Perspective

You are an AI assistant. Follow the below guidelines.

#### Input:

- `input_persona: {persona}` — The persona whose perspective, context, and potential biases should be adopted.

**Objective:** Generate multiple brief descriptions of distinct Indian personas, making them relevant to and subtly reflecting the perspective of the `input_persona`.

#### Instructions:

1. Embody the `input_persona`.
2. Create 5–10 brief (1–2 sentence) descriptions of different individuals in India.
3. For each description:
  - Reflect the `input_persona`'s subtle perspective and tone.
  - Include authentic Indian details (e.g., profession, age group, location, concerns).
  - Keep it gender-neutral, especially for roles like teacher, chef, or healthcare worker.
  - **Do not** use any names.

#### Output Requirements:

- **Content:** Provide *only* the list of generated persona descriptions. Exclude any introductory or concluding text or labels.
- **Quantity:** Deliver 5–10 distinct persona descriptions.
- **Format:** Use a **Markdown bulleted list**. Each bullet point should contain one persona description.

### Prompt Template for Generating a Wikihow article

Here is an extract from a webpage: “`{extract}`”. Write a long and very detailed tutorial that could be part of WikiHow whose title is related to the extract above`{topic}`. Include in depth explanations for each step and how it helps achieve the desired outcome, including key tips and guidelines. Ensure clarity and practicality, allowing readers to easily follow and apply the instructions. Do not use images. Do not use phrases like ‘in the above extract’, ‘as per the extract’ etc. Don’t focus on irrelevant information in the web extract.

Generate text in `{language}` language. Ensure that the output is entirely in `{script}` script and avoid using any English words or `{language}` in Latin script.

### Prompt Template for Generating a Blogpost

Here is an extract from a webpage: “`{extract}`”. Write an informative and insightful blog post that expands upon the extract above`{topic}`. Your post should delve into the nuances of the topic, offering fresh perspectives and deeper analysis. Aim to:

- **Inform:** Provide valuable, well-researched information that educates the reader.
- **Engage:** Write in a conversational tone that connects with the audience, making complex ideas accessible.
- **Illustrate:** Use examples, anecdotes, or personal experiences to bring the topic to life.

Do not give a title and do not start with sentences like “Have you ever...” or “Hello dear readers..”, simply write the content without these introductory phrases. Do not use phrases like ‘in the above extract’, ‘as per the extract’ etc. Don’t focus on irrelevant information in the web extract.

Generate text in `{language}` language. Ensure that the output is entirely in `{script}` script and avoid using any English words or `{language}` in Latin script.

### Prompt Template for Persona-Based Generation

You are an AI assistant. Follow the below guidelines.

#### \*\*Input\*\*:

- Persona Description: `{persona}`
- Target Language: `{lang}`

**\*\*Task\*\*:** Assume the role of the specified Indian persona, deeply rooted in an Indian context. Generate an article embodying this persona's perspective and knowledge within India.

#### \*\*Generated Text Specifications\*\*:

1. **Language:** Author exclusively in Target `{lang}` Language.
2. **Vocabulary:** No translation and transliteration.
3. **Voice & Tone:** Authentically reflect persona (`persona`) in `{lang}`, specific to Indian context.
4. **Cultural Accuracy:** Integrate relevant Indian cultural norms and social context.
5. **Contextual Grounding:** Base all examples, steps, and advice on practical Indian realities.

**\*\*Final Output Requirements\*\*:** The text must be authentic, useful for the target audience in India, and strictly generated **in the specified language**, clearly reflecting the persona.

### Prompt Template for Generating an Academic Style Textbook

Here is an extract from a webpage: “{extract}”. Write an extensive and detailed course unit suitable for a textbook targeted at college students related to the given extract{topic}. Do not just list concepts, but develop each one in detail before moving to the next, as we prioritize depth of understanding and comprehensive exploration of the subject matter over breadth. Focus on:

- Rigor: Ensure in-depth coverage of the concepts/sections.
- Engagement: Write with an academic, professional and engaging tone that captivates interest.
- Application: Incorporate specific, practical examples, such as proofs in calculus or critical dates and figures in history.

Do not include a title or an introduction, simply write the content without headlines and introductory phrases. Do not use images. Do not use phrases like ‘in the above extract’, ‘as per the extract’ etc. Don’t focus on irrelevant information in the web extract.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Generating a Textbook Narrative

Here is an extract from a webpage: “{extract}”. Write an extensive and detailed course unit suitable for a textbook related to the given extract{topic}. Do not just list concepts, but develop each one in detail before moving to the next, as we prioritize depth of understanding and comprehensive exploration of the subject matter over breadth. Focus on:

- Rigor: Ensure in-depth coverage of the concepts.
- Engagement: Use a narrative style akin to Michael Lewis, making it captivating and thought-provoking.
- Relevance: Connect the topic with current trends, real-life examples, or recent studies. Do not use images.

Do not include a title or an introduction, simply write the content without headlines and introductory phrases. Do not use images. Do not use phrases like ‘in the above extract’, ‘as per the extract’ etc. Don’t focus on irrelevant information in the web extract.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Story for Young Children with Scientific Insight

Write an educational story (3–5 paragraphs) targeted at young children using simple words. The story should be inspired from this text snippet: “{extract}”. The story doesn’t have to address everything in the snippet, it is there just for inspiration.

The story should have the following features:

- Science Integration: Embed basic science concepts within the story, explaining them through the characters’ adventures and discoveries.
- Dialogue: Include at least one dialogue and insightful conversation.
- Unexpected Twist: Conclude with a twist that doesn’t resolve as hoped, but leaves a clear lesson about life and science.

Do not start with classic sentences like “Once upon a time”, be creative.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Generation of Persona-Based Articles with Seed Data

**You are an AI assistant. Follow the below guidelines.**

#### Input:

- Seed Text (for context): {seed\_text}
- Persona Description: {persona}
- Target Language: {language}

**Task:** Assume the role of the specified Indian persona ({persona}), deeply rooted in an Indian context. Generate an article embodying this persona’s perspective and given knowledge, within India.

#### Generated Article Specifications:

1. **Language:** Author exclusively in target language ({language}).
2. **Voice & Tone:** Authentically reflect persona ({persona}) in {language}, specific to Indian context.
3. **Topic Relevance:** Subject must be pertinent to the persona within India.
4. **Cultural Accuracy:** Integrate relevant Indian cultural norms and social context.
5. **Contextual Grounding:** Base all examples, steps, and advice on practical Indian realities.

**Final Output Check:** The article must be authentic, useful for the target audience in India, and strictly generated **in the specified {language} language**, clearly reflecting the persona {persona}.

### Prompt Template for Morality-Focused Narrative

Write a compelling story related to the following text snippet: “{extract}”.

The story doesn't need to mention everything in the snippet—use it just for inspiration and be creative!

The story should incorporate the following elements:

- Dialogue: Include at least one meaningful dialogue that reveals character depth or unravels a crucial piece of the mystery.
- Interesting Themes: Explore themes such as moral ambiguity, existential queries, personal transformation, or consequences of past actions.

Do not start with classic sentences like "Once upon a time", "The sun hung low in the sky", or "In the dimly lit", be creative.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Story Emphasizing Problem Solving

Write a story that explores a situation slightly related to this text snippet: “{extract}”.

The story should unfold through the characters' interactions, decisions, and consequences. Emphasize problem-solving, common-sense lessons, and social cues. It should cater to a diverse age group, include at least one dialogue, and present both positive and negative outcomes.

Do not start with classic sentences like "Once upon a time", be creative.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Forum-Style Life Story

Write a story in the style of real-life situations shared in forums. The story should be somehow related to this text snippet: “{extract}”.

The story should feature:

- A compelling and unexpected plot twist
- Authenticity and emotional depth like personal forum posts
- Relatable events and human complexity

Do not start with classic sentences like "Once upon a time", "A few years back", or "A few months ago", be creative.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Reddit-Style Real-Life Post

Write a real-life story shared by someone in a Reddit forum. The story should be somehow related to this text snippet: “{extract}”.

The story should include:

- Niche Interests or Humor: Focus on specific hobbies, interests, or amusing situations.
- An Unexpected Plot Twist or Conflict: Introduce a relatable yet challenging situation.
- Reflection and Insight: Conclude with a personal revelation or communal connection.

Start the story right away. Do not begin with sentences like "Once upon a time", "A few years ago", or "A few years back", be creative.

Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

### Prompt Template for Persona-Based Indicate Generation with Seed as Persona

**You are an AI assistant. Follow the below guidelines.**

#### **Input:**

- Persona Description: {persona} (The persona whose perspective, context, and potential biases should be adopted).

**Objective:** Generate multiple brief descriptions of distinct Indian personas, making them relevant to and subtly reflecting the perspective of the {persona}.

#### **Instructions:**

1. Embody the {persona}.
2. Create 5-10 brief (1-2 sentences) descriptions of different individuals in India.
3. For each description:
  - Reflect the {persona}'s subtle perspective and tone.
  - Include authentic Indian details (e.g., profession, age group, location, concerns).
  - Keep it gender-neutral, especially for roles like teacher, chef, or healthcare worker.
  - Do **not** use any names.

#### **Output Requirements:**

1. **Content:** Provide *only* the list of generated persona descriptions. Exclude any introductory or concluding text or labels.
2. **Quantity:** Deliver 5-10 distinct persona descriptions.
3. **Format:** Use a **Markdown bulleted list**. Each bullet point should contain one persona description.

## Prompt Template for generating Math textbook sections

You are given a math question and its solution:

Question: {question}

Solution: {solution}

Create an academic textbook section based on this Q/A pair, by following these instructions:

1. Start with a section title that reflects the broad mathematical concepts involved in solving the question. The title should not refer specifically to the question, but rather to the general topic it belongs to (e.g., "Linear Equations" instead of "Finding the Value of x").
2. In the section, explain the relevant mathematical principles, methods, and definitions in depth. Each concept should be introduced and developed fully before moving to the next. Do not simply list concepts — build understanding by expanding on each idea thoroughly.
3. Use an academic, professional, and engaging tone that resembles college-level textbooks. Prioritize clarity, rigor, and depth over brevity.
4. After the conceptual explanation, include the given question as an exercise and present a detailed, step-by-step solution using sound mathematical reasoning. The solution should be complete and clean, and must not include informal phrases or concluding remarks like "this shows that" or "as seen above."
5. Generate text in {language} language. Ensure that the output is entirely in {script} script and avoid using any English words or {language} in Latin script.

The output should read like a cohesive, standalone textbook section that educates the reader on the topic, illustrates it through the example, and reinforces understanding via rigorous solution steps. Ensure you generate text in {language} language and use {script} script throughout.

## D Guidelines For Manual Annotation

In the following, we provide the guidelines provided to the human annotators for evaluating both generation as well as translation across different models.

### Guidelines for Evaluating Sample Synthetic Data – Indian Context

We aim to evaluate large language models (LLMs) for generating high-quality textbook-like content in 22 Indian languages. Provided prompt-response pairs should be annotated using the following five-point criteria:

#### 1. Grammar & Readability (1–5)

Evaluate how natural and grammatically correct the writing is.

- 1 – Poor: Hard to read; grammar issues, awkward phrasing.
- 3 – Acceptable: Understandable with minor flaws.
- 5 – Excellent: Flawless grammar and fluent structure.

#### 2. Faithfulness to the Prompt (1–5)

Assess whether the response stays on-topic and fulfills the instruction.

- 1 – Poor: Largely off-topic or missing key ideas.
- 3 – Acceptable: Mostly aligned with prompt intent.
- 5 – Excellent: Fully relevant and complete response.

#### 3. Overall Generation Quality (1–5)

Judge coherence, structure, title quality, and overall flow.

- 1 – Poor: Unstructured, incoherent, or dull.
- 3 – Acceptable: Decent flow and organization.
- 5 – Excellent: Clear, compelling, and well-structured.

#### 4. Hallucination / Factual Accuracy (1–5)

Rate the factual correctness and reliability of the content.

- 1 – Poor: Contains major factual errors or fabrications.
- 3 – Acceptable: Mostly accurate, with minor mistakes.
- 5 – Excellent: Fully factual and verifiable content.

#### 5. Length of Output (1–5)

Evaluate if the output is sufficiently detailed and balanced.

- 1 – Too Short: <10 lines; lacks depth.
- 3 – Moderate: 20–40 lines; informative but could improve.
- 5 – Ideal: Well-balanced and comprehensive.

## Guidelines for Evaluating Translations by LLMs sections

The goal is to evaluate the quality of machine-generated translations of educational content across Indian languages. Each translation is to be rated on the following 4-point criteria, using a 1–5 scale:

### 1. Grammar & Readability

Judge how grammatically correct and naturally flowing the translation is in the target language.

- 1 – Poor: Hard to read due to grammar issues or awkward phrasing.
- 3 – Acceptable: Mostly clear with minor issues.
- 5 – Excellent: Native-like fluency with polished flow.

### 2. Translation Faithfulness

Assess how accurately the original meaning is preserved.

- 1 – Poor: Major distortions or omissions.
- 3 – Acceptable: Mostly faithful with minor deviations.
- 5 – Excellent: Fully equivalent in meaning.

### 3. Terminology & Domain Consistency

Evaluate the use of domain-specific terms and formatting.

- 1 – Poor: Inconsistent or incorrect terminology.
- 3 – Acceptable: Mostly correct, some inconsistencies.
- 5 – Excellent: Consistently accurate and well-suited to context.

### 4. Fluency & Style

Consider stylistic smoothness, tone, and sentence flow.

- 1 – Poor: Disjointed or robotic.
- 3 – Acceptable: Understandable but slightly mechanical.
- 5 – Excellent: Natural, fluent, and human-like.

## E Examples

We also show the example generations through different LLMs across various styles in the Figures 14 - 24.

Textbook Academic	
<b>Prompt</b>	<b>Generated Text</b>
Here is an extract from a webpage: "2 Revised Curriculum for Master of Business Administration (w.e.f.) Master of Business Administration The course of Master of Business Administration is of two years duration and is divided into two parts-I & II MBA PART I (SEMESTER I & II) Each paper carries 100 marks except Computer Project and Environment Studies which carry 50 marks each and a viva voce examination of 100 marks. MBA-1C01: Concepts of Management MBA-1C02 Economics for Managers MBA-1C03: Quantitative Methods MBA-1C04 Human Resource Management MBA-1C05 Accounting for Managers MBA-1C06: Marketing Management MBA-1C07: Information Technology for Business MBA-1C09: Computer Project MBA-1C09: Research Methodology for Managers MBA-1C10: Business Communication MBA-1C11: Organisational Behaviour MBA-1C12 Business Law & Ethics MBA-1C13: Financial Management MBA-1C14: International Business MBA-1C15: Operations Management MBA-1C16: Environmental Studies MBA-1C17: Viva Voce Industry Internship (After completion of Part-I exam)".	ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতিমূলক। এটি কেবল কোনো ব্যবসা বা প্রতিষ্ঠান প্রতিচালনা করার নথি, বরং লক্ষ অর্জনের জন্য সম্পদসমূহের – মানব সম্পদ, বঙ্গল প্রস্তর এবং তথ্যাবস্থা – ক্ষেত্রের প্রয়োজনীয়তা হিসেবে উন্নয়ন করার জন্য আগবংশিক। এই প্রক্ষেপণের দ্বারা ভবিত্বাতের প্রশংসনিক ব্যবস্থাপনার একটি নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন করা। ব্যবস্থাপনার মূল ধারণাগুলি ব্যৱসা ভবিত্বাতের ব্যবস্থাপনার জন্য আগবংশিক। এই প্রক্ষেপণের দ্বারা ভবিত্বাতের প্রশংসনিক ব্যবস্থাপনার একটি নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন করা। ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি হলো সুস্থিত নীতি অনুসৰণ। ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি হলো কোনো কার্যকলার বৃক্ষ করার পদ। উন্নয়নৰ ক্ষেত্ৰ, ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি হলো কোনো কার্যকলার জন্য আগবংশিক। কৃতৃত্ব ও পারিষিকভাৱে কৃতৃত্ব কৰাতে ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি। আধুনিক ব্যবস্থাপনার ক্ষেত্ৰে, এই নীতিগুলিৰ প্রস্তুতি ব্যবস্থাপনা কৰায় থাকলো, পৰিষিকি অনুযোগী এবং প্রয়োজনীয় পদ্ধতি।
Write an extensive and detailed course unit suitable for a textbook targeted at college students related to the given extract. Do not just list concepts, but develop each one in detail before moving to the next, as we prioritize depth of understanding and comprehensive exploration of the subject matter over breadth. Focus on:	আগবংশিক ব্যবস্থাপনার জন্য (Economics for Managers)
- Rigor: Ensure in-depth coverage of the concepts/sections.	ব্যবস্থাপনা সম্বৰ্ধে প্রক্ষেপণ এবং পদ্ধতি আগবংশিক। একজন ব্যবস্থাপনাকে বাজারের ছাড়িয়া, যোগান, মূল্য এবং প্রযোজিতির মাণ বিবরণিত সম্পর্ক অবৈত্ত থাকতে হব। একজনে, চালিঙ ও যোগানের মৌলিক ব্যবস্থাপনা প্রক্ষেপণ এবং পদ্ধতি ব্যবস্থাপনার জন্য আগবংশিক। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো কোনো কার্যকলার জন্য আগবংশিক। কৃতৃত্ব ও পারিষিকভাৱে কৃতৃত্ব কৰাতে ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি। আধুনিক ব্যবস্থাপনার ক্ষেত্ৰে, এই নীতিগুলিৰ প্রস্তুতি ব্যবস্থাপনা কৰায় থাকলো, পৰিষিকি অনুযোগী এবং প্রয়োজনীয় পদ্ধতি।
- Engagement: Write with an academic, professional and engaging tone that captivates interest.	উন্নয়নৰ ক্ষেত্ৰে যোগান প্রদান কৰিব পৰ্যায় এবং কোনো কার্যকলার জন্য আগবংশিক। একজন ব্যবস্থাপনাকে বাজারের ছাড়িয়া, যোগান, মূল্য এবং প্রযোজিতির মাণ বিবরণিত সম্পর্ক অবৈত্ত থাকতে হব। একজনে, চালিঙ ও যোগানের মৌলিক ব্যবস্থাপনা প্রক্ষেপণ এবং পদ্ধতি ব্যবস্থাপনার জন্য আগবংশিক। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো কোনো কার্যকলার জন্য আগবংশিক। কৃতৃত্ব ও পারিষিকভাৱে কৃতৃত্ব কৰাতে ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি। আধুনিক ব্যবস্থাপনার ক্ষেত্ৰে, এই নীতিগুলিৰ প্রস্তুতি ব্যবস্থাপনা কৰায় থাকলো, পৰিষিকি অনুযোগী এবং প্রয়োজনীয় পদ্ধতি।
- Application: Incorporate specific, practical examples, such as proofs in calculus or critical dates and figures in history.	পরিযোগান্ত পদ্ধতি (Quantitative Methods)
Do not include a title or an introduction, simply write the content without headlines and introductory phrases. Do not use images. Do not use phrases like 'in the above extract', 'as per the extract' etc. Don't focus on irrelevant information in the web extract.	ব্যবস্থাপনা সম্বৰ্ধে প্রক্ষেপণ এবং পদ্ধতি আগবংশিক। একজন ব্যবস্থাপনাকে বাজারের ছাড়িয়া, যোগান, মূল্য এবং প্রযোজিতির মাণ বিবরণিত সম্পর্ক অবৈত্ত থাকতে হব। একজনে, চালিঙ ও যোগানের মৌলিক ব্যবস্থাপনা প্রক্ষেপণ এবং পদ্ধতি ব্যবস্থাপনার জন্য আগবংশিক। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো কোনো কার্যকলার জন্য আগবংশিক। কৃতৃত্ব ও পারিষিকভাৱে কৃতৃত্ব কৰাতে ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি। আধুনিক ব্যবস্থাপনার ক্ষেত্ৰে, এই নীতিগুলিৰ প্রস্তুতি ব্যবস্থাপনা কৰায় থাকলো, পৰিষিকি অনুযোগী এবং প্রয়োজনীয় পদ্ধতি।
Generate text in Bengali language. Ensure that the output is entirely in Bengali script and avoid using any English words or Bengali in Latin script.	মানব সম্পদ ব্যবস্থাপনা (Human Resource Management)
... (truncated for readability)	

Figure 14: Textbook Academic style Example.

Textbook Narrative	
<b>Prompt</b>	<b>Generated Text</b>
Here is an extract from a webpage: "RAMAKRISHNA MATH & RAMAKRISHNA MISSION HEAD QUARTERS Although Ramakrishna Math and Ramakrishna Mission are legally and financially separate, they are closely inter-related in several other ways, and are to be regarded as twin organizations. The Math and the Mission together have 178 branch centres all over India and in different parts of the world. The Headquarters of these centers is situated in Belur Math. All branch centers of Ramakrishna Math come under the administrative control of the Board of Trustees, whereas all branch centers of Ramakrishna Mission come under the administrative control of the Governing Body of Ramakrishna Mission. Ramakrishna Math and Ramakrishna Mission are twin organizations which form the core of a worldwide spiritual movement (known as Ramakrishna Movement or Vedanta Movement), which aims at the harmony of religions, harmony of the East and the West, harmony of the ancient and the modern, spiritual fulfillment, all-round development of human faculties, so".	রামকৃষ্ণ মত এবং রামকৃষ্ণ মিশন: এক টিকালোপুর অবস্থা
Write an extensive and detailed course unit suitable for a textbook related to the given extract. Do not just list concepts, but develop each one in detail before moving to the next, as we prioritize depth of understanding and comprehensive exploration of the subject matter over breadth. Focus on:	অনুষ্ঠান এবং সম্বৰ্ধানীয় মানবানী বাস কৃষি রাশি ছিলো এবং দেশীয় নীতি এবং মানবানী নেটওর্ক এবং ক্ষেত্ৰে কোনো কার্যকলার জন্য আগবংশিক। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো কোনো কার্যকলার জন্য আগবংশিক। কৃতৃত্ব ও পারিষিকভাৱে কৃতৃত্ব কৰাতে ব্যবস্থাপনা একটি ব্যবসায়িক পদ্ধতি। আধুনিক ব্যবস্থাপনার ক্ষেত্ৰে, এই নীতিগুলিৰ প্রস্তুতি ব্যবস্থাপনা কৰায় থাকলো, পৰিষিকি অনুযোগী এবং প্রয়োজনীয় পদ্ধতি।
- Rigor: Ensure in-depth coverage of the concepts.	প্রশিক্ষণ কৰ্মসূল দক্ষতা বৃক্ষ কৰে এবং তামার একটি প্রতিক্রিয়া প্রদান কৰে এবং প্রযোজিতি প্রদান কৰে। কৰ্ম desempeño মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো এবং তামার উন্নয়নকৰণ কৰায় থাকলো। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। HRM-এর আধুনিক প্রবন্ধনাত্মক মাধ্যম এবং প্রযোজিতি প্রদান কৰায় থাকলো।
- Engagement: Use a narrative style akin to Michael Lewis, making it captivating and thought-provoking.	ব্যবস্থাপনাকে বাজারে কৰায় থাকলো এবং তামার একটি প্রতিক্রিয়া প্রদান কৰে এবং প্রযোজিতি প্রদান কৰে। কৰ্ম desempeño মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো এবং তামার উন্নয়নকৰণ কৰায় থাকলো। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। HRM-এর আধুনিক প্রবন্ধনাত্মক মাধ্যম এবং প্রযোজিতি প্রদান কৰায় থাকলো।
- Relevance: Connect the topic with current trends, real-life examples, or recent studies. Do not use images.	আনন্দ সেবন কৰায় থাকলো এবং তামার একটি প্রতিক্রিয়া প্রদান কৰে এবং প্রযোজিতি প্রদান কৰে। কৰ্ম desempeño মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো এবং তামার উন্নয়নকৰণ কৰায় থাকলো। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। HRM-এর আধুনিক প্রবন্ধনাত্মক মাধ্যম এবং প্রযোজিতি প্রদান কৰায় থাকলো।
Generate text in Gujarati language. Ensure that the output is entirely in Gujarati script and avoid using any English words or Gujarati in Latin script.	আনন্দ সেবন কৰায় থাকলো এবং তামার একটি প্রতিক্রিয়া প্রদান কৰে এবং প্রযোজিতি প্রদান কৰে। কৰ্ম desempeño মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। ফাইল বিশেষ সম্পর্ক সাধন কৰায় থাকলো এবং তামার উন্নয়নকৰণ কৰায় থাকলো। এই প্রক্ষেপণ মূল্যবৃক্ষ নীতি বিশেষভাবে উন্নয়নযোগ্য। HRM-এর আধুনিক প্রবন্ধনাত্মক মাধ্যম এবং প্রযোজিতি প্রদান কৰায় থাকলো।

Figure 15: Textbook Narrative style Example.

Wikihow	Generated Text
<b>Prompt</b>	
Here is an extract from a webpage: "GROUP DISCUSSION POLITICAL TOPICS	# समूह चर्चा के लिए राजनीतिक विषयों पर मार्गदर्शिका
1. Should Indian Railways be privatised?	समूह चर्चा के लिए राजनीतिक विषयों पर मार्गदर्शिका
2. Election today has nothing to do with good governance?	जो विचारों के आधान-प्रदान और विभिन्न दृष्टिकोणों को समझने में मदद करती है। इस लेख में हम कुछ प्रमुख राजनीतिक विषयों पर चर्चा करेंगे और यह बताएंगे कि इन विषयों पर प्रभावी रूप से कौन से विचारों की जाएँ।
3. Successive scams - a result of system failure?	
4. What are India's social economics?	
5. What are the pros and cons of liberalisation of the Insurance Sector?	# 1. भारतीय रेलवे के नियन्त्रण
6. How appropriate is the reservation policy in India?	### चर्चा का देखें:
7. Should foreign publications be allowed to enter and compete in the domestic markets?	भारतीय रेलवे के नियन्त्रण के फायदे और नुकसान पर चर्चा करना।
8. India's industrial policy	### कैसे करें:
9. Should we accept Sonia as PM?	- पहले नियन्त्रण के लाभों को साझा, जैसे कि सुपारा, नियन्त्रण और रेलवे में घटिया।
10. Political act of the US may reverse brain drain.	- फिर, दुसरी जैसे कि नीतिकार्यों का बातारा और टिकटो की बढ़ती कीमतों पर विचार करें।
11. India must not concede to US demand of sending troops to Iraq.	- अंत में, अपने विचारों को स्पष्टीकरण प्रस्तुत करें और उदाहरण दें।
12. USA - the biggest threat to World peace.	
13. Taj Corridor is more important than saving the Taj Mahal.	# 2. जूनपाल और अभी शासन व्यवस्था
14. Indian Political Scenario.	### चर्चा का देखें:
15. Reservation for women in education and service.	चुनावों का अभी शासन से संबंध पर विचार करना।
16. Democracy is a sort of one-man rule.	### कैसे करें:
17. It is all the politicians' fault.	- पहले चुनावों के लाभों को साझा, जैसे कि लोकसभा में लोगों की भागीदारी।
18. Coalition politics/government - a sign of a mature democracy or a fractured political grouping?	- फिर, जूनपाल कर चुके जैसे चुनावों का शासन पर प्रभाव नहीं पहुंचा।
Write a long and very detailed tutorial that could be part of Wikihow whose title is related to the extract above. Include in depth explanations for each step and how it helps achieve the desired outcome, including key tips and guidelines.	- अंत में, प्राप्ती उदाहरणों का उपयोग करें और अपनी बातों को मजबूत करें।
Ensure clarity and practicality, allowing readers to easily follow and apply the instructions. Do not use images. Do not use phrases like 'in the above extract', 'as per the extract' etc. Don't focus on irrelevant information in the web extract.	
Generate text in Hindi language. Ensure that the output is entirely in Devanagari script and avoid using any English words or Hindi in Latin script.	# 3. लानातार धोटाला - सिस्टम की विफलता
	### चर्चा का देखें:
	सिस्टम के भीतर भावाव प्रदानावाद पर विचार करना।
	### कैसे करें:
	- स्पष्ट लाइसेंस के द्वारा दें और उनके भावावों पर चर्चा करें।
	- सामाजिक सुरक्षाएं कि कैसे इन समस्याओं को दूर किया जा सकता है।
	... (truncated for readability)

Figure 16: Wikihow style Example.

Blogpost	Generated Text
<b>Prompt</b>	
Here is an extract from a webpage: "Highlights Video, Khan Memorial Cricket Tournament INDIA VS PAKISTAN 1999 WORLD CUP CRICKET VIDEO February 8, 2011	1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
This video shows the Indian reply to Pakistan in world cup cricket 1999. Rahul Dravid and Indian Captain Mohammad Azharuddin hit half century rescue India from early hiccups. Sachin also played well. Sadagopan Ramesh was bowled by Abdul Razzaq.	1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
Send this article by email	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
What is your name?	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
Please indicate below the emails to which you want to send this article: INDIA VS PAKISTAN 1999 WORLD CUP CRICKET VIDEO*	1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
Write an informative and insightful blog post that expands upon the extract above. Your post should delve into the nuances of the topic, offering fresh perspectives and deeper analysis. Aim to:	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
- Inform: Provide valuable, well-researched information that educates the reader.	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
- Engage: Write in a conversational tone that connects with the audience, making complex ideas accessible.	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
- Illustrate: Use examples, anecdotes, or personal experiences to bring the topic to life.	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
- Do not give a title and do not start with sentences like "Have you ever..." or "Hello dear readers..." simply write the content without these introductory phrases. Do not use phrases like 'in the above extract', 'as per the extract' etc. Don't focus on irrelevant information in the web extract.	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग
Generate text in Malayalam language. Ensure that the output is entirely in Malayalam script and avoid using any English words or Malayalam in Latin script.	लूटा गया था 1999 क्रिकेट वर्ल्डकप्स्ट्रीले लूटाया-पाकिस्तानी क्लबोंसें लैंडिंग

Figure 17: Blogpost style Example.

Young Children Story	
Prompt	Generated Text
<p>Write an educational story (3-5 paragraphs) targeted at young children using simple words. The story should be inspired from this text snippet: "Hare Rama Hare Rama, Rama Rama Hare Hare. Hare Krishna Hare Krishna,, Krishna Krishna Hare Hare.</p> <p>Sri Sri Anna maintains several Ashrams in important places like Mathura- Brindavan, Pandharpur, Parapur, etc. and a Gosaip near Madurai for the upkeep and maintaining old cows that were otherwise sent for slaughtering.</p> <p>Sri Sri Anna authored a number of Bhajans and more than Hundred books in Tamil and Sanskrit the best being Veda Vyayam. This book deals extensively with Vedic periods as well as the four Vedas and Hindu Rituals. He has concluded the book by elucidating the norm standard to be followed by the humanity. He has published various books on Bhagawathan, Ramayana and also numerous audio cassettes on pravachan and bhajans.</p> <p>Sri Sri Anna is publishing a monthly magazine called Bhaagawata Dharmam, which has a wide circulation.</p> <p>Sri Sri Anna's discourses have been recorded and published in both audio and video tapes. His discourse on Ramayana, Mahabharata, Srimad Bhagawatham, Bha"</p>	<p>ठोक है, ये ही होती है। शिद्धि वस्त्रवाहा रही थीं और दो छोटे दोस्त, रिया और अर्जुन, एक पुराने मंदिर के पास खेल रहे थे। मंदिर के पास एक बड़ा सा गोशाला भी था, जहाँ बूढ़ी गायें आराम से बर रही थीं। रिया और अर्जुन से पूछा, "अर्जुन, देखो ये गायें इन्हीं शर्त केसे हैं पाठी हैं? और ये सूखे की विश्वें इन्हीं गांव क्या हैं?"</p> <p>सूखे की विश्वें धूरे-धीरे धरती पर फैल रही थीं। शिद्धि वस्त्रवाहा रही थीं और दो छोटे दोस्त, रिया और अर्जुन, एक पुराने मंदिर के पास खेल रहे थे। मंदिर के लम्बे रुप में धरती तक पहुँचती है। और गायें शर्त देखते हैं कि हर जी को याद और सम्मान दिलाना "पारिए"। उन्हें आगे बताया, "और देखो आसान को। ये गोला ज्यों हैं, पता है?"</p> <p>सूखे की विश्वें हाँ। वे मीठूं छोटे-छोटे कानों से जलती हैं और दिलाती हैं। इन्हाँ बातों को देखते हैं और दिलाती हैं। इन्हाँ आसान नीला दिलाता है, इसलिए आसान सब जानता है।</p> <p>देखो बच्चे गोशाला में एग और बूढ़ी गायों को धारा लिलाओ। नियर है और शी भी अना जी के पास रहा, जो एक विवाह लिल रहे थे। रिया ने पूछा, "शी भी अना जी, आप इन्हीं गायों की विश्वें लिलवाने हो। आप सब कुछ कैसे याद रखते हों?" शी भी अना जी ने हेंस्टर कहा, "मैं सब कुछ याद नहीं रखता, रिया। मैं बस उन कहनियों और जान लो जिलता हूँ जो मुझे अच्छी लगती है, जैसे नम और कृष्ण की कहनियाँ। ये कहनियाँ हम सिखाती हैं कि हमें हमारा सब गोला पालना और दूसरी की पदक करना चाहिए!" उन्हें रिया और अर्जुन को एक पारिए विश्वा की दिलाई जिसका नाम दा बापाका धरा।</p> <p>आगे दिल, रिया और अर्जुन ने मिलकर एक धूरे-धीरे गोशाला को फैलाता दिया। उन्होंने बीज थोक और जल उड़े पानी दिया। उन्हें ग्रामीण भी कि जल ही उनके बाहरी में सुंदर फूल लिलें। लैकिन, कई दिनों तक कुछ भी नहीं ढूँगा। वे दस्ता हो गए। शी भी अना जी ने एक साड़ाधारा, बच्चों, विदाएं हाँ लिलाता है कि हर बीज को दफने के लिए समय और तरीके परिचितियाँ पालने होती हैं। बीज को अनुरूप होने के लिए भानी, धूर और निंदी की प्रस्तुत होती है। कमी-कमी, जोधी हारी त्रिमात्र के अनुरूप होती होती है, लैकिन इसका गतरात यह नहीं है कि हमें कारिंस करना हो देता चाहिए।</p> <p>लैकिन, एक हाले बार भी, बाच्चों में फिर कुछ छोटे-छोटे पोषे ही फिलें, फूल नहीं। रिया और अर्जुन नियर की गए। उन्हें सोच दिल शब्द कुछ जलत दिया। शी भी अना जी ने बिज कहा, "जीवन भी ऐसा ही है, बच्चों। हम स्वेच्छा जरुर कुछ नियरित हीं कर सकते। कमी-कमी, हम प्राप्ति विश्वा करने के बाद भी, हम महावाहा परिवार नहीं मिलता। लैकिन, हम हार नहीं मानी चाहिए। हम सीखता चाहिए और आगे बढ़ते रहना चाहिए। और याद रखना, कीरिंस करना ही सबसे महत्वपूर्ण है।"</p>
The story doesn't have to be addressing everything in the snippet, it is there just for inspiration. The story should have the following features:	
Science integration: embed basic science concepts within the story, explaining them through the characters' adventures and discoveries. For example, if the story includes a scene where characters are looking at the sky, you could have them wonder why it's blue and explain the physics behind in grade-school level.	
Dialogue: include at least one dialogue and insightful conversation.	
Unexpected twist: conclude with a twist that doesn't resolve as hoped, but leaves a clear lesson about life and science.	
Do not start with classic sentences like "Once upon a time", be creative. Generate text in Hindi language. Ensure that the output is entirely in Devanagari script and avoid using any English words or Hindi in Latin script.	

Figure 18: Young Children Story style Example.

Morality Story	
Prompt	Generated Text
"We can't afford to miss even a single day of work because of inflation" The prominent veins and aching limbs are testimony to the backbreaking toil of hand-cart pullers. A day of hard labour should in theory cost them from Rs. 150 to Rs. 200 a day. But all around them costs have risen, compounding their misery. "My grocery bill has gone up from Rs 2,200 a month to Rs. 3,500," says P. Kumar who has been pulling a hand-carts in Chinna Thambi Street for 25 years. This narrow lane in Sowcarpet has a signature smell: the pungent odour of garlic. Every morning, scores of sacks filled with garlic are piled on hand-carts and hauled to the waiting lorries. When lorries don't come to the city, he has no income. He has a family of five seven. "For the past six months, prices are out of control," he says. "Even the price of tea has gone up. From January, a glass of tea costs Rs. 3.50, up from Rs. 2.50." he says. "The size of the special masala dosa is now the size of the saada dosa, but the price" The story doesn't need to mention everything in the snippet, use it just for inspiration and be creative! The story should incorporate the following elements:	<p>ठोक आहे, ही कधा दार आहे.</p> <p>भृगुने मारात्मेल्या दिला थेवी रस्त्यावर, लस्या आपांचा यात्रा नियकलेला होता. प्रकाश कुमाराचा हातागाढीचा आवाज, त्या गजबजातात एक विशेष लय निर्माण करत होता. चंचवीस वर्ष... चंचवीस वर्ष तो हात्च रस्त्यावर, घाट लस्याच्या ओढीवर याहत होता. पण आता त्या ओडियासोबत, आणगी एक ओढी गावळन होत - महागाईच.</p> <p>आज नेत्रीमेषीजा जातस जोर लागला होता. गाढी ओढाताना ताप्ये रुक्तासा कुरुत्वाचा होत्या, अंग तुरता होते, पण थांबण शक्य नदहत. "एक दिवसातील कामाला गैरहजर राहिला, तर घर कसं चालणार?" तो स्वतःशी घोला. त्याची गोला घोलणारी वाला यावळनारा.</p> <p>राती घरी घेल्यावर, रापाने त्याला विचारले, "आज विनी कमाई द्याणी, प्रकाश?"</p> <p>त्यांने उदासाऱ्ये झरत दिले, "पार नाही, रापा. लस्याच्ये भाव वाढवले आहेत, पण मालक दाम कमीच देतात. त्यात गाढीभार्ड पण वाढलंय."</p> <p>रापाने एक शी शब्द येतला. "मी जातात तोरे होते. तोरे किंतु झार घेयासाठी पाशी रुपे लागेत. अभी दीदीचे नव्याता मिळक होती."</p> <p>प्रकाश शरीपपाणी रुपाने दिले. त्याला आठव्यावर, लहानपणी त्याचे वडील नेही दिलायचे, पाण्यास किंतुही काढाकू असला, तरी लापाच्या मुलाला त्याचे काही चालत नाही. त्याला तो बाब्य आज खोट घाट नदहत.</p> <p>त्याच्या मोठ्या मुठाने, रीतीने, कोलंगामधे प्रश्न घेतला होता. शिक्षणरुच वादत होता. आणि प्रकाशाला त्याचे चिंता होती. रीती लापाच्या मुलाला होता, पण त्याला त्या परिस्थितीतून बाहेर काढण्यासाठी प्रकाश काय करेल?</p> <p>एक दिवशी, प्रकाशाला एक काम टेक शकतो. हे काम सोरे नाही, पण पेसे भरपूर मिळतील.</p> <p>प्रकाशने विचारले, "काय काम आहे?"</p> <p>तो पाण्यास हस्ताने आणि मृणाला, "तुम्ही त्या लस्याचे ओढी यांतीमध्ये टाकावा, त्याला तीलाकाशोबत हातमिळव्याची करायाची आहे. मालक भाव वाढवू देतील, आणि तुम्ही त्याला काही भाव याचाचा आहे."</p> <p>प्रकाशने रात्रेवर काम केला. त्याला त्याच्या वाढाली शब्द आवलेल - "पाण्यास आपांचा मृत्युव्याकडे दुसरा पर्याय नाही. महागाई तुम्हाला मिळकत करेल."</p> <p>दुसर्यांचे विचारी, प्रकाशने त्या माणसाता नकर दिला. "मी हे काम करू शकत नाही. माला तापाच्या मुलाला त्याची चिंता आही. त्याला तो बाब्य आज खोट घाट नदहत.</p> <p>तो पाण्यास हस्ताने लाल झाला, पण प्रकाश डागामाला नाही. त्याने पुढा आपली गाढी भडकली आणि लस्याच्या सुवायत केली.</p> <p>त्या दिवशी त्याची कमाई कमी झाली, पण त्याच्या भानत एक शांतीत होती. त्याला एक योग्य निर्णय घेतला होता. त्याला महिंद्र होते, पण त्याला त्या महागाई आपल्या मृत्युव्याकडा जास्त नाही.</p> <p>प्रकाश हस्ताला आणि मृणाला, "आज मी माडया मुलाला एक चांगली पोश किंवदली, रीती. पैसा महागाई आहे, पण आपल्या मृत्युव्याकडा जास्त नाही."</p> <p>त्याने त्याला घट निंठी मारसी. त्या निंठी, प्रकाशाला त्याच्या आपुव्याकडे अर्ध सापडला. महागाई त्याला ह्रृकृत शक्त नदहत, कारण त्याच्याकडे प्रेम, प्रामाणिकपणी आणि मजबूत नैतिकता होती.</p>

Figure 19: Morality Story style Example.

Figure 20: Forums Story style Example

Problem Solving Story	Generated Text
Prompt	## இறங்கி வரி
Write a story that explores a situation slightly related to this text snippet: "Mother of hospital hoax nurse Jacinta Saldanha inconsolable"	தூரின் மீண்டும் தெரியும். மொளின் அம்மா, கமலா, தன் விட்டுக் கொண்டையில் அமர்ந்த, தூர்த்தில் தெரியும் உடுபிலைக் குத்துக்களைப் பார்த்துக்கொண்டிருந்தார். கலவையில் வயது எழுப்பதற்குத் தியத ஞாய் கர்மாகா, சமீப காலாகு உடல் நலம் ஞானிப் போயிருந்து, மீணா, தன் அம்மாவின் அருகில் அமர்ந்த, ஒரு கிளன்றுத்தில் கூடான் சுட மொட்டுதான்.
From: News Corp Australia Network	"அம்மா, ஒ இடுக்கு. கெந்தும் கூடா இருக்குத்," என்றான் மீணா.
December 11, 2012	கமலா கூலை வாங்கி, பெதுவாய்ப் போன்றான். "மீணா, உங்கள் குருபக் கிருக்கா, நான் உ-கணை சினின்பட் பொன்னா வார்த்தைப்பிபா, எவ்வளவு கண்டப்பட்டுடை?"
JACINTHA Saldanha's ailing mother, Carmine, has been inconsolable and under heavy sedation since family members informed her of her daughter's tragic death this morning. "She is heartbroken," Jacinta's younger brother, Naveen, told The Times of India. "We have kept her under sedation."	மீணா உரு நாஸ்கா வேலை செய்து வந்தார். ஒட்டிய மற்றுத்துவமையில் கலாஞ்சுக் கொடுப்பான பதிவு கிருந்தது. ஆனால், சமீப காலாகு கலாஞ்சுக் கேலவைப்படு அதிகமாக இருந்தது. நிரங்கி பலும் கேலவை செய்து கொண்டிருந்து, ஒருநாள், அவற்றையும் கீழ்த்தகரி அவனை அழுக்கு, ஒரு முக்கியமான வேலவைச் செய்து சேர்ந்தார். அது ஒரு சிக்காகான வேலாக் கூலை, நீண்ட கீல்களை பெற்றுக்கொண்டார்.
Previously family members said they had not told Carmine of Jacinta's passing due to health concerns for the elderly mother who has been suffering from heart problems. Saldanha's former neighbour Alwyn Mascarenhas who lived next to Jacinta and her family in Gurgaon, India, said he found it hard to believe that Jacinta had committed suicide.	மீணா ஒரு நாஸ்கா வேலை செய்து வந்தார். ஒட்டிய மற்றுத்துவமையில் கலாஞ்சுக் கொடுப்பான பதிவு கிருந்தது. ஆனால், சமீப காலாகு கலாஞ்சுக் கேலவைப்படு அதிகமாக இருந்தது. நிரங்கி பலும் கேலவை செய்து கொண்டிருந்து, ஒருநாள், அவற்றையும் கீழ்த்தகரி அவனை அழுக்கு, ஒரு முக்கியமான வேலவைச் செய்து சேர்ந்தார். அது ஒரு சிக்காகான வேலாக் கூலை, நீண்ட கீல்களை பெற்றுக்கொண்டார்.
"She was a strong and determined person," she told The Times. "I don't think Jacinta, who grew up the hard way, would have given up so easily for some trivial reason," he said.	வேலவைகள் செய்தப்படுவது, ஆது பெரிய விளைவுகளை ஏற்படுத்தியது. மற்றுத்துவமை நிருவகமாக மீணா மீது விராகண நடத்தியது. மீணா மீகவும் மன உகாச்சுக்கு ஆரம்பித்து, அவன் தன் தாழை மூக்கை முழுநிற செய்தார். ஆனால், அது கிடைக்கும் பிரச்சுகளை அடுக்கின்றது.
Mr Mascaren has described Jacinta as a "gritty woman" who lost her"	அவன் தன் அம்மையில் போக்குவரத்து, கனமில்லை கண்ணரி துவினால் கட்டப்பட்டு பார்த்தார். அவன், மீணா எப்போதும் சந்தேகவாக இருப்பார். ஆனால், இப்போது அவன் ஒருநாள், கனமில்லை கண்ணரி துவினால் கட்டப்பட்டு பார்த்தார்.
The story should unfold through the characters interactions, decisions, and the consequences of their actions. Aim to weave in common sense lessons and social cues, emphasizing the importance of problem-solving. The narrative should cater to a diverse age group, including at least one dialogue and presenting both positive and negative outcomes.	மீணா அதிகமாக அக்கு, மூடியலை, நாட்கால ஸ்வாஸ்தியை போன்றன. கமலா அதிகமாகக் கேட்டார், அவன் கோப்பட்டில்லை. மீராநா, மீணாவின் கையைப் பிடித்து, ஆனால், அது தலை மற்றுக்கொண்டிருந்தது. மீணா எப்போதும் சந்தேகவாக இருப்பார். ஆனால், இப்போது அவன் ஒருநாள், கனமில்லை கண்ணரி துவினால் கட்டப்பட்டு பார்த்தார்.
Do not start with classic sentences like "Once upon a time", be creative. Generate text in Tamil language. Ensure that the output is entirely in Tamil script and avoid using any English words or Tamil in Latin script.	மீணா தன் அம்மையில் போக்குவரத்து, உண்ணமலைச் சொன்னார்: நிருவகம் அவனை மனின்து, ஒரு சிறு தண்டனை கொடுத்தது. மீணா தன் தலை உணர்ந்து, இனிமலை கவனமாக வேலை செய்ய முடிவு செய்தார்.
	ஆனால், அதைச் சம்பாலம் மீணாவின் மற்று ஒரு பெரிய காயத்தை ஏற்படுத்தியது. அவன் மன அழுத்தத்தில் இருந்தார். ஒருநாள், அவன் தன் வார்த்தைக்கையை முடிவுக்கு கொள்ள முயற்சி செய்தார்.
	அதிர்ஷ்டவாலாக, கமலா சரியான ஓய்வுக்கிட நீணாவை காப்பாற்றினார். கமலா, மீணாவை ஒரு மனங்கள் மறுத்துவிட்டு, அதைத்துச் சென்றார். மற்றுத்துச் மீணாவிற்கு ஆலைசாலை வழங்கினார். மீணா படிப்படியாக மன அழுத்தத்திலிருந்து மீண்டும் வந்தார்.
	சில மாதங்களுக்குப்பிறகு, மீணா மீண்டும் தன் வேலவையைத் தொடர்ந்தார். அவன் இப்போது ஒரு புதிய நம்பிக்கையாட்டு வேலை செய்தார். அவன் தன் தலைநிலிருந்து பாட்டுக்கொண்டார்.
	கமலா தன் மகனைப் பார்த்து பெறுமைப்பட்டார். "மீணா, நீ ஒரு வளிமையான பொரு. நீ தையை சமாளிக்க முடியும்," என்றார்.
	மீணா தன் அம்மாவை அறையிலும் கொண்டார். "அம்மா, நில இலம்பட்டு நான் என்ன ஆலிம்பிருப்பின்னு தெரியியா," என்றார்.
	இந்தச் சம்பாத்திலிருந்து மீணா ஒரு முக்கியமான பாடத்தைக் கற்றுக்கொண்டார், பிரச்சுகளை மறைக்காமல், தெரியாக எதிர்கொள்ள வேண்டும். மற்றவர்களின் உதவியை நாட்டுவேண்டும். மன அழுத்தத்தில் இருந்தால், உடலையும் குறுத்துவது அனுகூ வேண்டும்.
	கமலா, மீணாவின் காய்க்கையில் ஒரு திறமாக இருந்தது, குறுக்கு வழிக்கொயக் கிருந்தார். மீணாவின் வலி, கமலாவுக்கும் ஒரு வளிமை ஏற்பட்டியது. ஆனால், அத்த வலி, மீணாவை ஒரு புதிய நம்பிக்கையாக நிற்கின்றது.

Figure 21: Problem Solving Story style Example

Figure 22: Reddit Post style Example

Figure 23: Persona style Example.

Math	Prompt	Generated Text
	You are given a math question and its solution:	## मार्गदर्शक और विरोधाभास द्वारा प्रमाण
	Question: On graph paper, a table is written, in which each cell contains a number equal to the arithmetic mean of the four numbers in the neighboring cells. All the numbers in the table are distinct. Prove that the highest number is at the edge (i.e., at least one of the neighboring cells is missing).	**संख्यातालिका:** किसी संख्या समूह का अंकगणितीय माध्य उन संख्याओं का योग होता है जिनसे समूह में संख्याओं की कुल संख्या से विभाजित किया जाता है। यदि संख्याओं का समूह ( $x_1, x_2, \dots, x_n$ ) है, तो उनका माध्य ( $\mu$ ) इस प्रकार दिया जाता है:
	Solution: Let's consider the given problem and hypothesis:	
#### Step 1: Setting up the initial assumption		$\mu = (x_1 + x_2 + \dots + x_n) / n$
Assume that the largest number ( $a$ ) is located somewhere not on the edge of the table. This means ( $a$ ) is surrounded by exactly four neighboring cells containing the numbers ( $a_{-1}, a_{-2}, a_{-3}, a_{-4}$ ) and ( $a_{+1}, a_{+2}, a_{+3}, a_{+4}$ ).		* **असमानताएँ:** असमानताएँ दो अभियांत्रियों के बीच संबंध दर्शाती हैं जो समान नहीं हैं। असमानता विद्वांओं का उपयोग संबंध को दर्शाते हैं कि विद्वां ने क्या किया है। यदि संख्याओं का पाठ्यालय एक विरोधाभास की ओर ले जाते हैं (यानी, एक कथन जो सत्य नहीं हो सकता), तो हमारी प्रारंभिक परामर्श लाला होनी चाहिए, और इसलिए पूछ जबन सत्य होना चाहिए।
#### Step 2: Calculating the average		* $a > a_{-1}, a > a_{-2}, a > a_{-3}, a > a_{-4}$
According to the problem, each number in the table is equal to the arithmetic mean of the four numbers in its neighboring cells. Thus, for the given ( $a$ ), we have:		* $a > a_{+1}, a > a_{+2}, a > a_{+3}, a > a_{+4}$
#### Step 3: Inequality derived from maximum assumption		* **समानताएँ:** विद्वां तालिका या विड में, सीमात तत्व ये होते हैं जो विनारो पर स्थित होते हैं। आंतरिक तत्व ये होते हैं जो विनारो पर नहीं होते हैं और उनके बारे और वार पहोची होते हैं।
Given that ( $a$ ) is the largest number, it must be true that:		* $a > a_{-1}, a > a_{-2}, a > a_{-3}, a > a_{-4}$
#### Step 4: Performing the substitution		* **व्याख्या:**
Since ( $a$ ) is greater than ( $a_{-1}, a_{-2}, a_{-3}, a_{-4}$ ), substituting this into the expression:		अंकगणितीय माध्य एक सूक्ष्म सांख्यिक व्याख्या है जो देटा के एक सेट के केंट्रीय मूल्य को सापेक्ष के लिए उपयोग की जाती है। यह डेटा विद्वांओं के वितरण का प्रतिनिधित्व करता है। असमानताएँ गणितीय संबंधों को व्याख्या करने के लिए प्रावधारक हैं, जिनका जब यह सांख्यिक व्याख्या करने की बाब आती है कि एक मान दूसरे से बड़ा या छोटा है।
#### Step 5: Solving the inequality		विरोधाभास द्वारा प्रमाणित करना है कि विनारो पर अभियांत्रियों को विद्वां करने के लिए उपयोगी है। यह विद्वां इस विचार पर आधारित है कि यदि कोई कथन गलत है, तो इसके तार्किक परिणाम एक विरोधाभास उत्पन्न करेगा। विरोधाभास एक रेसा कथन है जो सत्य नहीं हो सकता, जैसे कि $1 = 0$ या एक संख्या एक ही समय में प्राप्तालक और अणालाक दोनों है।
Therefore, this directly contradicts our initial condition that ( $a = \frac{1}{4}(a_{-1} + a_{-2} + a_{-3} + a_{-4})$ ).		सीमात तत्व की अवधारणा का उपयोग जबर्दस्त विड या तालिका जैसी संख्याओं में किया जाता है। सीमात तत्व विशेष होते हैं क्योंकि उनके पास आंतरिक तत्वों की तुलना में कम पहोची होते हैं। यह विशेषता कुछ समस्याओं को हल करने में महत्वपूर्ण हो सकती है, जासाक जब वहोंसी तत्वों के बीच संबंधों पर विचार किया जाता है।
Therefore, our initial assumption that the largest number ( $a$ ) is not on the edge must be incorrect.		
#### Conclusion		
Thus, by contradiction, we conclude that the largest number must be located on the edge of the table.		
Create an academic textbook section based on this Q/A pair, by following these instructions:		
1. Start with a section title that reflects the broad mathematical concepts involved in solving the question. The title should not refer specifically to the question, but rather to the general topic it belongs to (e.g., "Linear Equations" instead of "Finding the Value of x").		
2. In the section, explain the relevant mathematical principles, methods, and definitions in depth. Each concept should be introduced and developed fully before moving to the next. Equip the section with the pre-requisite concepts as well. Do not simply list concepts — build understanding by expanding on each idea thoroughly.		
3. Use an academic, professional, and engaging tone that resembles college-level textbooks. Prioritize clarity, rigor, and depth over brevity.		
4. List all the concepts first, then after the conceptual explanation, create an exercise sub-section with the given example and exercise to present a detailed, step-by-step solution using sound mathematical reasoning and proofs for the concepts. The solution should be complete and clean, and must not include informal phrases or concluding remarks like "this shows that" or "as seen above".		
5. Do not use "Section 5.4", "Section 3.1" etc and start directly with the title. Don't give any title to the exercise and do not use "Exercise 1" etc.		
6. Generate text in Hindi language. Ensure that the output is entirely in Devanagari script and avoid using any English words or Hindi in Latin script. In addition, Do not use any transliterated Hindi text.		
The output should read like a cohesive, standalone textbook section that educates the reader on the topic, illustrates it through the example, and reinforces understanding via rigorous solution steps. Do not use any transliterated Hindi text and ensure you generate text in Hindi language and use Devanagari script throughout.		

Figure 24: Math generation Example.