MILA (MULTILINGUAL INDIC LANGUAGE ARCHIVE): A DATASET FOR EQUITABLE MULTILINGUAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are predominantly trained on high-resource languages such as English, leaving low-resource languages marginalized. This imbalance is particularly acute in India, where most Indic languages lack clean, large-scale digitized corpora despite having hundreds of millions of speakers. Building equitable representation for these languages requires not only greater data volume, but also novel pipelines for acquisition, curation, and validation. We present ***MILA***, the largest curated Indic multilingual dataset to date, spanning 7.5 trillion tokens across 16 of India's 22 official languages. The dataset is constructed through a multi-stage process combining large-scale crawling, OCR pipelines tailored to Indic scripts, LLM post-corrected translations, synthetic augmentation via the **Indic-Persona Hub**, data distillation, and rigorous filtering. Each stage is validated by expert linguists, ensuring both linguistic fidelity and cultural authenticity. Alongside, we release **Indic-MMLU**, a translation and verification of MMLU into 16 Indian languages, providing the first large-scale multilingual benchmark for evaluation. We further introduce multiple *general* and *domain-specific taxonomies* (finance, Ayurveda, agriculture, and law) to enable creation of targeted pre-training and post-training corpora for Indic use cases. To assess the impact of these resources, we conduct extensive experiments across translation pipelines, OCR incorporation, synthetic supervised fine-tuning (SFT) data generation, and continual pretraining analyses. Across all tasks, models trained on MILA demonstrate stronger performance on Indic-MMLU and achieve improved parity with English, underscoring the value of curated pipelines for equitable multilingual modeling. By bridging resource gaps at scale and validating through language experts and **synthetic rewriting**, MILA promises to be a foundational archive for inclusive large-scale language modeling in the Indic context. All resources, including the MILA dataset, Indic-MMLU benchmark, and accompanying taxonomies, are released in an anonymous GitHub repository for reproducibility and community use.[1]

1 INTRODUCTION

The progression of language models from early monolingual architectures to advanced multilingual systems highlights the expanding capabilities of NLP. Early models such as BERT (Devlin et al. (2018)) and GPT (Radford & Narasimhan (2018); Radford et al. (2019); Brown et al. (2020)), built upon the Transformer architecture, demonstrated the effectiveness of self attention in text understanding, but in very limited languages. The multilingual generation of models, including mT5 (Xue et al. (2020)), XLM R (Conneau et al. (2019)), Bloom (Muennighoff et al. (2023)), LLaMA (Touvron et al. (2023a;b); Grattafiori et al. (2024)), Gemma (Team et al. (2024a;b; 2025)), Mistral (Jiang et al. (2023)), Qwen (Bai et al. (2023); Yang et al. (2024); Qwen et al. (2025); et al. (2025)), and Nemotron (Nvidia et al. (2024)), has broadened the scope of NLP by enabling understanding and generation across numerous languages. Large scale datasets such as Common Crawl (Common Crawl Foundation), Wikipedia (Wikimedia Foundation) dumps, CCMatrix Schwenk et al. (2019), mc4 (Xue et al. (2020)), OSCAR (Ortiz Suárez et al. (2019)) and Dolma Soldaini et al. (2024), provide the backbone for training such multilingual language models.

[1] https://github.com/anonymous-submitter0104/iclr-submission

However, the availability of high quality data is highly uneven across languages. Even in these massive multilingual datasets, Indic languages remain severely underrepresented, despite being the native language of billions of people, making each token precious. This imbalance reflects the global disparity in digital resources, where languages with limited digital presence are systematically disadvantaged. To address this gap, we present the largest curated Indic multilingual dataset to date, MILA, a 7.5 trillion token corpus encompassing 16 languages. Instead of relying on a single source or method, the curation strategy combines diverse web sourced corpora, OCR recovery for under digitized scripts, translation pipelines and synthetic augmentation, and data distillation while refining quality with expert linguistic validation.

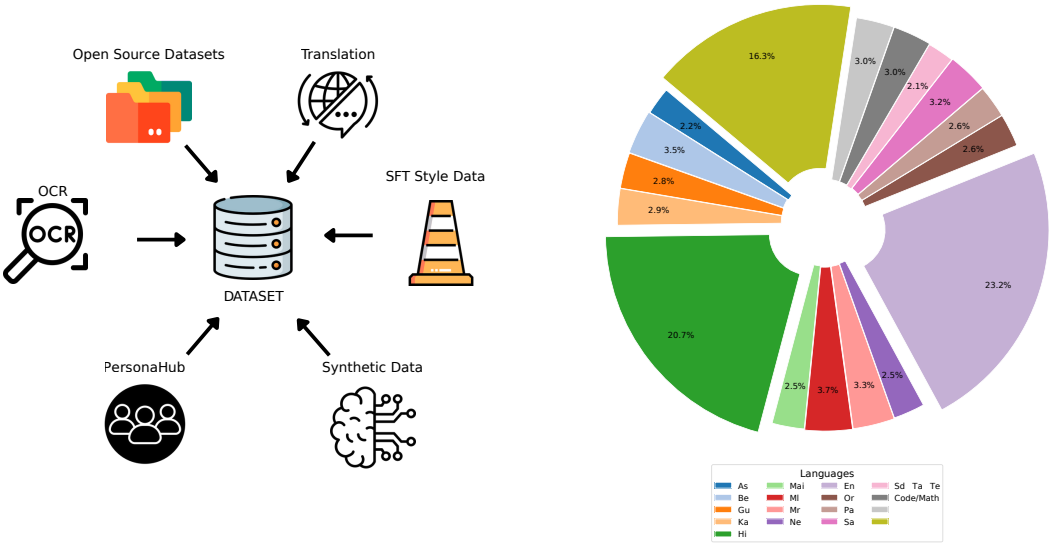

Figure 1: Dataset Sources and Distribution

Beyond the dataset itself, we conduct evaluation using the Indic MMLU (Hendrycks et al. (2020)) benchmark and introduce a parity based metric to assess fairness across languages. This metric quantifies performance disparities, providing a more equitable and interpretable measure of how well multilingual models serve low resource Indic languages. Practical insights are drawn from the pipeline, highlighting both the strengths and limitations of current OCR and translation techniques for Indic languages, and providing guidance for future dataset creation in low resource settings. Concrete algorithms for all experiments and pipelines are given in the Appendix.

2 Related Work

English heavy and multilingual corpora. Large scale corpora such as RedPajama (Weber et al. (2024)), SlimPajama (Shen et al. (2024)), DCLM (Li et al. (2025)), The Pile (Gao et al. (2021)), Zyda (Tokpanov et al. (2024b;a)), and TxT360 (Tang et al. (2024)) have driven major advances in NLP but remain heavily concentrated in English, leaving the majority of the world's languages underrepresented. Multilingual collections (mC4 (Xue et al. (2020)), OSCAR (Ortiz Suárez et al. (2019)), CC100, ROOTS (Laurençon et al. (2023)), ParaCrawl (Bañón et al. (2020)), FineWeb2 (Penedo et al. (2025)), CulturaX (Nguyen et al. (2023)), MultiUN (Eisele & Chen (2010)), Dolma (Soldaini et al. (2024))) broaden coverage, yet their per language depth is highly uneven; Indic languages, in particular, receive only fractional representation compared to English. This imbalance limits the cultural and linguistic grounding available to multilingual LLMs and reduces downstream performance on region specific tasks.

Indic focused datasets and their limitations. Several recent efforts target Indian languages, but important limitations persist. Parallel collections such as Samanantar (Ramesh et al. (2022)) and synthetic corpora like Sangraha Synthetic (Khan et al. (2024)) rely on translations or non-native sources, which can produce text lacking cultural authenticity. Monolingual efforts such as IndicCorp (Doddapaneni et al. (2023)) provide higher quality native text but remain modest in scale, and are

limited in script and dialect coverage. Taken together, prior datasets either offer scale without native depth, or native depth without scale, motivating the multi-stage curation pipeline we present here.

Current efforts in curation and augmentation. Beyond dataset scale, much of the recent work on multilingual corpora has focused on pipelines that convert raw web data into usable training material. Core components such as deduplication, robust language identification, and quality scoring (Lee et al. (2022); Khan et al. (2025); Zhang & Salle (2023); Sharma et al. (2024)) have been emphasized in large scale projects (Weber et al. (2024); Shen et al. (2024); Gao et al. (2021)), yet these methods are typically optimized for high resource languages and degrade significantly under the noisy, code mixed conditions common in Indic data (Ousidhoum et al. (2025)).

OCR remains a significant bottleneck for Indic languages such as Devanagari, Bengali, and Tamil. As highlighted by Mathew et al. (2024), specialist systems face higher error rates than for Latin scripts due to challenges including complex script segmentation, Akshara level modeling, font variations, and Unicode reordering. Augmentation and synthetic generation offer a complementary path, leveraging back translation, self training, and multilingual LLMs, but often lack cultural grounding, with quality varying widely across Indic language pairs (Ousidhoum et al. (2025); Yu et al. (2022)). Moreover, most existing resources suffer from limited domain coverage and stylistic diversity, with a heavy reliance on web text or translated material, which limits LLM generalization across literature, news, technical writing, and colloquial registers. Together, these efforts illustrate both the promise and the limitations of current curation practices, highlighting the need for pipelines that are explicitly adapted to the linguistic, script, domain, and cultural characteristics of Indic languages.

Evaluation and parity. Benchmarks such as FLORES (Goyal et al. (2022)), IndicGenBench (Singh et al. (2024)) and MILU (Verma et al. (2025)) attempt to measure cross lingual performance, but results consistently show wide gaps between English and Indic languages. Prior analyses often report absolute scores, but relatively fewer works study parity ratios across languages, i.e., how close low resource language performance is to English within the same model. This remains a crucial but underexplored metric for assessing equality of representation in multilingual LLMs.

3 METHOD

This section details the pipelines and data sources employed to construct our Dataset.

3.1 DATA ACQUISITION

Part of our corpus is incorporated from multiple sources similar to Pile Gao et al. (2021), RedPajama Weber et al. (2024), and C4 Raffel et al. (2023). This includes about **5 billion** words gathered through multi-source web crawling of multilingual websites, forums, and academic repositories, as well as over 1700 open datasets from HuggingFace[2]. To mitigate cultural and linguistic gaps found in translated or synthetic corpora (Yao et al. (2024)), we emphasize curated book collections, amounting to approximately **32 billion** words across 16 languages, sourced primarily from Archive.org[3] (around 1 million books across domains ranging from mathematics to agriculture) and the National Digital Library of India (NDLI)[4] (about 28,500 curriculum-aligned, licensed documents). These provide authentic content and curriculum focused materials that complement the breadth of crawled and opensource data. Collecting these books and academic papers posed challenges such as stalled pipelines, duplication, and week-long processing times (Brown et al. (2024)).

This framework allowed efficient ingestion of millions of multilingual and domain-diverse documents, ensuring the resulting corpus is clean, license-compliant, and immediately suitable for large-scale pretraining, particularly in Indic and curriculum-aligned contexts.

3.2 DATA CURATION

High-quality training data is essential for robust language models. Low-quality or misclassified text can degrade performance and introduce undesirable behavior. Prior works, including Paullada et al.

[2] https://huggingface.co
[3] https://archive.org
[4] https://ndl.iitkgp.ac.in

To tackle these, we developed optimized, source-specific pipelines and high-concurrency, asynchronous crawlers for Archive.org and structured ingestion for NDLI, that reduced processing from 7 days to 24 hours, tripled throughput, and lowered compute usage by 40%, while ensuring fault tolerance and full provenance tracking. A metadata-first strategy enabled early governance, quality control, and deduplication, using a canonical schema (identifiers like ISBN, DOI, archive IDs; bibliographic fields; technical attributes; license data; integrity checks) and URL/MD5-based collapsing of redundant copies, which was critical for Archive.org's large-scale collections.

Table 1: Book Corpus By Language (Books, Pages and Word Counts)

Language	# PDFs	# Pages	Word Count
Hindi	396.12 K	7.53 M	4.15 B
Marathi	124.22 K	3.02 M	1.26 B
Malayalam	65.03 K	2.18 M	1.06 B
Telugu	77.86 K	5.93 M	1.53 B
Tamil	43.59 K	5.28 M	1.44 B
Kannada	41.71 K	4.08 M	1.01 B
Sanskrit	44.49 K	10.09 M	2.68 B
Bengali	41.25 K	10.95 M	3.10 B
Urdu	126.03 K	32.15 M	10.03 B
English	45.10 K	2.57 M	0.89 B
Total	**1 M**	**84.00 M**	**27.15 B**

(2021); Liu et al. (2024); Yu et al. (2024); Rae et al. (2022), emphasize the importance of auditing and curating datasets. However, existing curation pipelines, while effective for widely studied languages like English, are often insufficient for Indian languages. Each language presents unique challenges in morphology, script, and code-mixing, requiring specialized pipelines. To address this, we designed a multi-stage curation framework using NVIDIA NeMo Curator[5] and custom scripts tailored for multilingual Indian corpora. At the document level, we employ in-house model based quality classifiers (fastText) for all languages. High-quality documents comprising **3.7 trillion tokens** across all languages was retained as is, while medium- and low-quality documents were fed into a synthetic rewriting pipeline, producing high-quality text that preserves the linguistic characteristics of the original data, as detailed in Section 3.5.

Table 2: Benchmark Results: Conventional vs Curated

Model	ARC Challenge	ARC Easy	Hella Swag	Hella Swag Hi	MMLU	MMLU Hi
Conventional	46.5	73.6	73.5	28.9	41.3	26.2
Curated	53.6	74.2	73.8	41.4	46.2	34.6

Our pipeline also incorporates standard filtering and cleaning steps. UnicodeReformatter ensures textual consistency, exact deduplication removes bitwise identical documents, and GPU-accelerated fuzzy deduplication eliminates near duplicates from template variations in large-scale web corpora (Lee et al. (2022); Khan et al. (2025)). Following Mendu et al. (Mendu et al. (2025)), sensitive and unsafe content is handled with a two-stage toxic filtering process: initial rule-based filters flag toxic documents, and multilingual model inference reclassifies false positives to preserve safe content (RoBERTa). Finally, PiiModifier detects and redacts personally identifiable information, including names, addresses, emails, and phone numbers. By combining language-specific quality classification, synthetic rewriting, and standard cleaning pipelines, we produce a clean, diverse, safe, and legally compliant dataset suitable for large-scale pretraining and fine-tuning. Benchmark results of a 2.9B parameter dense model trained on conventional vs curated data (2 Trillion Tokens En/Hi) is given in Table 2, showcasing the effectiveness of our curation pipeline in 2 languages.

3.3 OCR Pipeline

OCR formed a critical component of our dataset creation pipeline, directly addressing the scarcity of digitized content in Indic languages. While languages like Hindi and Tamil have partial digital presence, low-resource ones such as Maithili and Sindhi remain largely undocumented. To bridge this gap, we identified underrepresented languages, collected 5-6 million pages from print materials and scanned books, converting otherwise inaccessible texts into machine readable form. In the OCR pipeline, almost 37 percent of scanned books suffered from faded ink, irregular printing, and evolving orthographies, making raw scans hard to use. Preprocessing pipelines involving denoising, contrast enhancement, and binarization were necessary to obtain clean inputs for OCR. Yet, even

[5] https://github.com/NVIDIA-NeMo/NeMo

high-quality scans posed challenges: Indic scripts contain stacked ligatures, conjunct consonants, and diacritics that generic OCR systems fail to capture reliably. Diversity in scripts (Devanagari, Bengali, Tamil, Telugu, etc.) and variation in layouts (books, newspapers, manuscripts, tables) demanded tailored pipelines for line detection, layout analysis, and text normalization. To improve fidelity, OCR outputs underwent postprocessing through confidence based routing and language specific correction layers, repairing ligature breaks, spacing artifacts, and broken Unicode sequences. This iterative process highlighted the interdependence of data and OCR: robust OCR requires curated data, and high quality data creation depends on reliable OCR pipelines. To illustrate these challenges, we conducted a series of controlled experiments. Due to legal complications in releasing original scanned documents, we crafted the Indic-Synthetic-OCR-Benchmark-Small (ISOB-Small), a synthetic OCR benchmark spanning 22 Indian languages. By generating 110 synthetic pages with diverse layouts (multi-column, tables, figures, equations) and realistic degradations (blur, watermark, shadow, folds, font variation) encountered during the processing of documents, ISOB-Small provided a stress test that exposed how poorly generic OCR systems transfer to Indic scripts.

As shown in Table 3 (a), we compared OCR-only models against VLMs on both existing Indian OCR benchmarks and ISOB-Small, tracking performance across multiple dimensions. While VLMs offered broader coverage, they suffer from hallucination for Indic languages. Traditional OCR models also face problems but are easier to identify and resolve. Table 3 (b) presents results after applying our language specialized preprocessing and postcorrection. These targeted enhancements improved benchmark performance and were critical for creating high-quality, machine-readable Indic text for downstream NLP and language model training. We assessed the performance of a dense 310M model on both raw OCRed text and quality-enhanced text. Pretraining on raw OCRed data produced noisy loss and perplexity curves, reflecting the inconsistencies and challenges inherent in unprocessed, low resource Indic corpora. In contrast, training on processed text yielded smooth convergence, demonstrating that careful preprocessing stabilizes model training. To validate the effectiveness of our pipeline, we employed LLM-assisted quality checks as a judgment method, leveraging models such as GPT-OSS-120B (OpenAI et al. (2025)), Deepseek (DeepSeek-AI et al. (2025)), and Qwen (Bai et al. (2023); Yang et al. (2024); Qwen et al. (2025); et al. (2025)) to assess semantic consistency. The results are showcased in the Appendix.

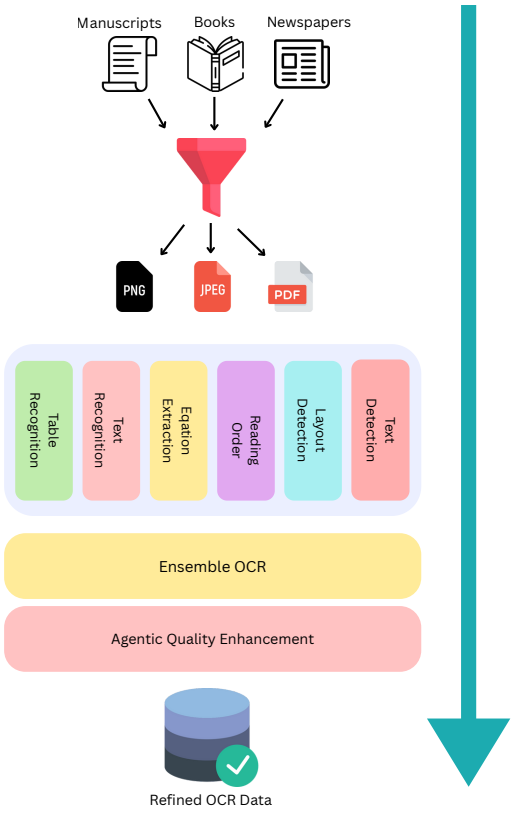

Figure 2: OCR Pipeline

3.4 Translation Pipeline

A key challenge in curating Indic datasets is the scarcity of high-quality monolingual and parallel corpora. Translation based data generation mitigates this gap by producing parallel corpora that enable cross-lingual transfer from resource rich languages to low resource Indic languages, a capability critical for multilingual large language models (LLMs). Moreover, as cited by Chen et al. (2023); Lin et al. (2024) downstream performance for math, STEM, and code gets a boost with the parsing of parallel corpora. However, as seen in Table 4 (a), no single translation model is universally optimal. Specialist machine translation systems excel in narrow contexts but often lack generalization, while generalist LLMs provide coverage but frequently struggle with fidelity in low-resource Indic

List of Models	Bhashini				Mozhi			
	CER	WER	PI-WER	Char3 gram F1	CER	WER	PI-WER	Char3 gram F1
Llama-4-Scout-17B-16E-Instruct	0.259	0.445	0.398	0.672	4.35	1.38	0.619	0.31
NuMarkdown-8B-Thinking	0.361	0.537	0.508	0.556	53.31	9.21	0.677	0.168
Llama-4-Maverick-17B-128E-Instruct_final	0.4	0.58	0.418	0.645	12	4	0.72	0.22
Qwen2.5-VL-72B-Instruct	0.676	0.847	0.45	0.613	18.22	4.16	0.677	0.266

(a) Without Preprocessing and Postprocessing

List of Models	CER	WER	PI-WER	Char3 gram F1	ISOB-Small (22 Languages,) 110 Pages WER
dots.OCR	0.168	0.253	0.23	0.801	0.8616
dots.OCR - postcorrected	0.085	0.145	0.12	0.91	0.8214
Surya	0.2	0.28	0.138	0.867	0.8982
Surya - postcorrected	0.095	0.16	0.11	0.925	0.8774

(b) With Preprocessing and Postprocessing

Table 3: Model Performance Benchmarks for Different Processing Pipelines

Evaluation including multilingual embeddings, back-translation, and vision, language similarity using CLIP/SIGLIP embeddings, confirmed improvements in semantic fidelity, even when conventional metrics like WER/CER couldn't. Validation plots for the 310M model are shown in Figure 3, and detailed observations and model architecture for these experiments are provided in the appendix, underscoring that processing digitized Indic corpora requires iterative refinement and tailored pipelines to support reliable downstream language modeling.

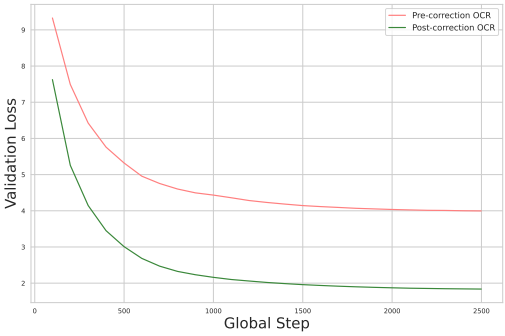

Figure 3: OCR Pipeline

languages. This tension underscores the need for carefully designed pipelines rather than reliance on any one system.

List of Models	Translation Benchmarks														
	As	Be	Gu	Ka	Hi	Mai	Ml	Mr	Ne	Or	Pa	Sa	Sd	Ta	Te
NLLB-200-3.3B	nan	48.58	52.10	56.87	52.67	44.08	49.35	47.03	45.93	46.05	49.48	25.28	52.49	54.37	48.24
NLLB-moe-54B	nan	49.86	53.30	57.03	53.08	46.63	51.47	47.85	45.10	45.34	48.75	25.56	53.46	55.72	48.71
hunyuan-mt	nan	42.22	41.38	46.62	45.01	8.40	1.91	43.02	0.83	0.95	1.04	18.80	42.94	40.03	39.75
deepseek v3.1 Think	39.10	44.30	47.95	53.12	47.56	39.59	46.86	44.80	47.23	44.34	46.80	25.37	48.90	48.68	46.05
Llama-4-Maverick-17B	40.35	47.09	48.71	53.21	47.57	42.03	44.88	46.64	44.65	39.34	45.88	28.13	48.05	47.42	57.34

(a) Translation Model Benchmarks Without Processing

List of Models	Translation Benchmarks														
	As	Be	Gu	Ka	Hi	Mai	Ml	Mr	Ne	Or	Pa	Sa	Sd	Ta	Te
IT2 Processed	48.40	51.71	55.53	58.71	54.97	49.49	55.99	51.00	56.01	52.28	51.08	30.24	56.86	58.65	50.88
IT2	45.10	48.67	53.38	55.62	52.26	47.04	52.23	49.33	53.42	50.47	49.82	27.70	53.03	54.77	49.19

(b) IndicTrans2 with Preprocessing and Postprocessing

Table 4: Translation Model Performance on FLORES Benchmarks using chrF++

Our translation pipeline begins with synthetic augmentation and translation from English and other resource-rich languages into 16 Indic languages. To improve quality and fidelity, we implement a robust LLM-based post-correction phase that repairs syntactic and semantic inconsistencies, enhances context preservation, and addresses subtle morphological and syntactic variations. Human evaluation is integrated at this stage, with 3 language evaluators reviewing initial outputs to guide model

selection and ensure culturally aligned linguistic representations. For long and complex content, particularly in mathematics, STEM, formal proofs, and code, we employ a chunking strategy, dividing sequences into manageable segments and providing summaries of preceding chunks as context, as shown in the appendix. This approach maintains dependencies across segments, preventing information loss in multi-step reasoning tasks, such as proofs or code execution. By combining chunking, autoregressive translation, and LLM-based post-correction, our pipeline generated coherent, high-fidelity translations across x billion tokens, significantly reducing errors and inconsistencies that would otherwise arise in single-pass approaches.

3.5 Synthetic Rewriting and Data Distillation

Data distillation is another critical step in addressing the scarcity of high-quality, culturally relevant datasets for Indian languages. While large language models and synthetic data pipelines provide vast quantities of text (DatologyAI et al. (2025); Patel et al. (2024)), much of this content originates from Western perspectives, reflecting predominantly English-centric knowledge, norms, and reasoning patterns. Translations of existing datasets, such as Sangraha Synthetic (Khan et al. (2024)), only partially address this gap; they may preserve linguistic fidelity but often fail to capture cultural context, local reasoning, and domain-specific nuances essential for Indian applications. Distillation provides cleaner and more efficient access to world knowledge from state-of-the-art models while offering domain control and flexibility in shaping the data. Since strong models for Indianness or Indian languages are lacking, leveraging existing SoTA LLMs allows us to embed Indian values and knowledge more effectively than through custom data construction pipelines. This approach ensures both data efficiency and controllable alignment with local cultural and domain-specific needs.

To implement this, we developed a multi-stage distillation framework centered on the creation and use of Indian virtual personas. All models used for this pipeline have been verified to have a relatively high performance in Indic MMLU as seen in Table **??**. Inspired by Ge et al. (2025), we built the **Indic PersonaHub**, a large-scale repository of over 300 million Indian personas spanning 1400+ domains, each designed to capture diverse Indian identities, values, and domain expertise. These personas form the backbone of our pipeline, guiding the generation of synthetic augmentation across tasks such as Chain-of-Thought reasoning, multi-turn dialogues, summaries, question-answer pairs, and cross-lingual transformations. We further leverage them to natively generate SFT-style instruction–response pairs, ensuring models are aligned not only linguistically but also culturally. By anchoring data generation in personas, the pipeline moves beyond simple translation or replication of Western corpora, yielding outputs that are contextually appropriate and authentically Indian.

To operationalize the persona-driven framework, we adopt a dual methodology for generating supervised fine-tuning datasets in the Indian context. The first component leverages personas to generate synthetic long-form articles, multi-turn dialogues, and reasoning-focused responses that embed cultural, historical, and societal nuances grounded in Indian context. This approach ensures that the synthetic data reflects not only linguistic fidelity but also Indian modes of reasoning and domain expertise. The second component focuses on systematically transforming unstructured and semi-structured sources, ranging from OCRed documents and transcripts to web scrapes, into high-quality question–answer datasets. Raw text is segmented into coherent chunks, validated for cultural relevance, and classified into domains such as Healthcare, Finance, History, and Culture. From these validated chunks, we generate self-contained questions with both concise and detailed answers, balancing factual precision with contextual depth for both fine tuning and generation of safety datasets. Together, these two pipelines yield a diverse instruction-focused corpus that is simultaneously grounded in real Indian knowledge sources and expanded through persona-driven synthesis.

3.6 Data Organization

Building a multilingual Indic dataset of 7.5T tokens requires governance far beyond conventional pipelines, as emphasized by Gebru et al. (2021) and Jernite et al. (2022). Unlike English corpora, Indic data faces fragmentation, noisy digitization, and heterogeneity across a dozen+ scripts. Sources span textbooks, newspapers, and social media with inconsistent annotation and licensing. Without strict governance, this diversity degrades reproducibility, fairness, and compliance. The challenge is not only scale but control: ensuring low-resource languages are preserved in the long tail, Unicode normalization does not collapse distinct scripts, and every transformation remains auditable.

Without rigorous governance and taxonomy, a trillion-token Indic dataset risks becoming brittle and unusable. To address this, we implement a governance-first AI data lakehouse unifying storage, lineage, metadata, governance, and versioning at petabyte scale. Complementing this, spanning 1,400+ domains, is our **Taxonomy**, which provides consistent structure and coverage across tasks; a representative subset is included in the appendix.

Scalable Storage. Resilient 3FS + MinIO storage with Raw, Curated, and Feature Store zones.
Lineage Tracking. OpenLineage + Marquez standardize lineage across Spark, Airflow, and Kafka.
Metadata Cataloging. DataHub organizes trillions of tokens into a searchable knowledge graph.
Governance. Apache Ranger and OPA enforce both broad compliance and fine-grained policies.
Versioning. Delta Lake and DVC ensure reproducibility and preservation of long-tail Indic corpora.

3.7 Linguist Validation

A critical component of building a high-quality indic multilingual dataset was a rigorous human-in-the-loop linguistic validation applied across all pipelines, including OCR, synthetic data generation, translation, and data distillation. Native language experts and linguists evaluated outputs iteratively on multiple dimensions: fluency, adequacy, grammar, tone, vocabulary richness, cultural appropriateness, and readability. Low quality outputs were flagged, corrected, and reintegrated, with pipelines rerun until consistently high scores were achieved, ensuring that each language and task leveraged the most effective, specialized pipeline.

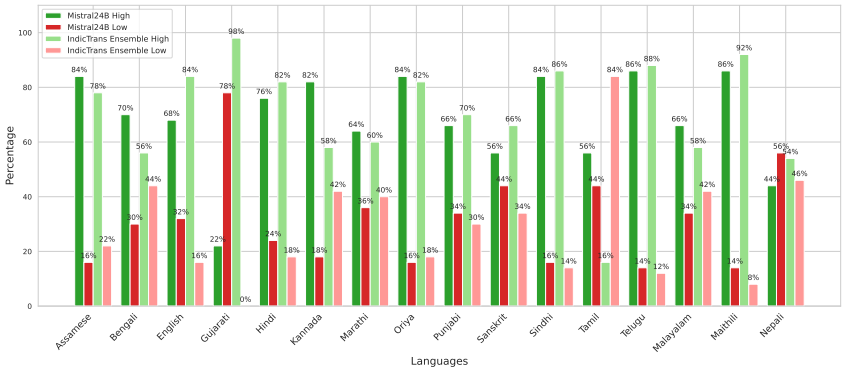

Figure 4: Readability Comparison: Mistral-214B vs IndicTrans2 NLLB Ensemble

To quantify this process, we compared two translation models, Mistral24B (Jiang et al. (2023)) and IndicTrans2 (Khan et al. (2024)) and NLLB (Team et al. (2022)) ensemble, using readability as a representative metric (Figure 4). Mistral-24B-Instruct (Jiang et al. (2023)) excelled in languages like Assamese (84%), Bengali (70%), and Hindi (76%), while ensemble of IndicTrans2 (Khan et al. (2024)) and NLLB (Team et al. (2022)) performed better in English (84%), Gujarati (98%), and Telugu (88%). Such comparisons guided model selection for each language, ensuring outputs were not only syntactically correct but also culturally and contextually aligned. This evaluation went beyond readability scores. For each task including OCR, translation, guided rewriting, and data distillation, we used a human-in-the-loop scoring system across syntactic correctness, readability/phrase structure, coherence, contextual inappropriateness, and domain- and information-level relevance. Multiple candidate pipelines were assessed per language, with the highest-scoring pipeline selected. Low-quality outputs were flagged, corrected, and reintegrated in an iterative refinement loop until all metrics consistently met high standards. This ensured each language and task leveraged a specialized, validated pipeline, preserving linguistic integrity, cultural context, and factual fidelity.

4 Experiments

To evaluate the effectiveness of MILA, we design a set of experiments that quantify both absolute performance and fairness of representation across Indic languages. As a first step, we take the Qwen3 600M (et al. (2025)) pretrained checkpoint and measure its Indic MMLU (Hendrycks et al. (2020)) score. We then continually pretrain this checkpoint on MILA and re-evaluate its Indic

MMLU performance. This direct before-and-after comparison highlights the impact of our dataset on improving reasoning and knowledge coverage for Indic languages.

Beyond absolute scores, we compute parity, defined as the ratio of a model's MMLU score in a given Indic language to its score in English. By measuring parity for both the original checkpoint and the continually pretrained checkpoint, we capture how fairness evolves during training. An increase in Indic parity demonstrates that our dataset not only improves raw performance but also promotes more balanced representation across languages.

$$\text{Parity}_L = \frac{\text{MMLU score in language } L}{\text{MMLU score in English}} \tag{1}$$

For completeness, we also evaluate Indic MMLU performance of several strong multilingual baselines, including mT5-XL (7B) Xue et al. (2020), BLOOMZ-7B Muennighoff et al. (2023), LLaMA-2-7B Touvron et al. (2023b), Gemma-7B (Team et al. (2024a;b; 2025)), Mixtral-7B Jiang et al. (2024), and Granite-7B Mishra et al. (2024). Results for these models are provided in the appendix. Together, these evaluations demonstrate both the effectiveness and fairness gains enabled by MILA.

5 RESULTS

Table 5 presents absolute scores for the original checkpoint and the continually pretrained checkpoint. The results demonstrate consistent gains across all Indic languages, indicating that exposure to MILA substantially improves reasoning and knowledge coverage in low-resource and high-resource languages alike.

Table 5: Indic MMLU Score across Indic languages for Qwen3-600M.

Model	As	Bn	En	Gu	Hi	Kn	Ml	Mr	Ne	Or	Pa	Sa	Sd	Ta	Te	Avg-Indic
qwen3-600M-original	0.2965	0.3020	0.3678	0.2950	0.3190	0.2906	0.2933	0.3002	0.2968	0.2861	0.2951	0.2968	0.2802	0.2962	0.2987	0.3012
qwen3-600M-cpt	0.3190	0.3270	0.3720	0.3180	0.3420	0.3130	0.3170	0.3240	0.3200	0.3090	0.3190	0.3200	0.3040	0.3200	0.3220	0.3250

To quantify fairness, we compute parity for both the original and pretrained checkpoints (Table 6). Parity captures the ratio of performance in each Indic language relative to English. We observe clear improvements in average Indic parity after continual pretraining, highlighting that MILA not only increases absolute performance but also promotes more balanced representation across languages. For completeness, we also evaluate Indic MMLU on strong multilingual baselines such as mT5-XL (7B), BLOOMZ-7B, LLaMA-2-7B, Gemma-7B, Mixtral-7B, and Granite-7B. These results are included in the appendix and provide additional context for the effectiveness of MILA in improving both performance and equitable coverage of Indian languages.

Table 6: Indic MMLU Parity for Qwen3-600M.

Model	As	Bn	Gu	Hi	Kn	Ml	Mr	Ne	Or	Pa	Sa	Sd	Ta	Te	Avg-Indic
qwen3-600M-original	0.806	0.821	0.802	0.867	0.791	0.797	0.816	0.807	0.778	0.802	0.807	0.762	0.806	0.813	0.819
qwen3-600M-cpt	0.857	0.879	0.855	0.919	0.841	0.852	0.871	0.860	0.830	0.857	0.860	0.817	0.860	0.865	0.874

6 CONCLUSION

In this work, we present a carefully curated Indic dataset to address the scarcity of high-quality training data for low-resource languages. Using Param-2.9B for ablation experiments Qwen3-600M for final experiments, we show that this dataset not only boosts absolute task performance but also improves parity across languages, as measured by Indic-MMLU. The dataset was constructed with attention to linguistic accuracy, diversity, and coverage across 16 Indic languages, reflecting the challenges of low-resource research. Our results highlight the central role of curated data in enabling large language models to perform fairly and robustly across diverse linguistic contexts, complementing advances in model scale and architecture.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417/.

Megan A. Brown, Andrew Gruen, Gabe Maldoff, Solomon Messing, Zeve Sanderson, and Michael Zimmer. Web scraping for research: Legal, ethical, institutional, and scientific considerations, 2024. URL https://arxiv.org/abs/2410.23432.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*, 2023.

Common Crawl Foundation. Common crawl dataset. https://commoncrawl.org/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

DatologyAI, :, Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Charvi Bannur, Christina Baek, Darren Teh, David Schwab, Haakon Mongstad, Haoli Yin, Josh Wills, Kaleigh Mentzer, Luke Merrick, Ricardo Monti, Rishabh Adiga, Siddharth Joshi, Spandan Das, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining, 2025. URL https://arxiv.org/abs/2508.10975.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang,

Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. 01 2010.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025. URL https://arxiv.org/abs/2406.20094.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2021. URL https://arxiv.org/abs/1803.09010.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 05 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00474. URL https://doi.org/10.1162/tacl_a_00474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,

Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,

Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, pp. 2206–2222. ACM, June 2022. doi: 10.1145/3531146.3534637. URL http://dx.doi.org/10.1145/3531146.3534637.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Arham Khan, Robert Underwood, Carlo Siebenschuh, Yadu Babuji, Aswathy Ajith, Kyle Hippe, Ozan Gokdemir, Alexander Brace, Kyle Chard, and Ian Foster. Lshbloom: Memory-efficient, extreme-scale document deduplication, 2025. URL https://arxiv.org/abs/2411.04257.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh Khapra. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15831–15879. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.843. URL http://dx.doi.org/10.18653/v1/2024.acl-long.843.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen,

Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023. URL https://arxiv.org/abs/2303.03915.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

Peiqin Lin, André FT Martins, and Hinrich Schütze. A recipe of parallel corpora exploitation for multilingual large language models. *arXiv preprint arXiv:2407.00436*, 2024.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*, 2024.

Minesh Mathew, Ajoy Mondal, and C. V. Jawahar. *Towards Deployable OCR Models for Indic Languages*, pp. 167–182. Springer Nature Switzerland, December 2024. ISBN 9783031784958. doi: 10.1007/978-3-031-78495-8_11. URL http://dx.doi.org/10.1007/978-3-031-78495-8_11.

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. Towards safer pretraining: Analyzing and filtering harmful content in webscale datasets for responsible llms, 2025. URL https://arxiv.org/abs/2505.02009.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. Granite code models: A family of open foundation models for code intelligence, 2024. URL https://arxiv.org/abs/2405.04324.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023. URL https://arxiv.org/abs/2309.09400.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340b technical report, 2024. URL https://arxiv.org/abs/2406.11704.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Nedjma Ousidhoum, Meriem Beloucif, and Saif M. Mohammad. Building better: Avoiding pitfalls in developing language resources when data is scarce, 2025. URL https://arxiv.org/abs/2410.12691.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. Datadreamer: A tool for synthetic data generation and reproducible llm workflows, 2024. URL https://arxiv.org/abs/2402.10379.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025. URL https://arxiv.org/abs/2506.20920.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022. URL https://arxiv.org/abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. Ccmatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944, 2019. URL http://arxiv.org/abs/1911.04944.

Vasu Sharma, Karthik Padthe, Newsha Ardalani, Kushal Tirumala, Russell Howes, Hu Xu, Po-Yao Huang, Shang-Wen Li, Armen Aghajanyan, Gargi Ghosh, and Luke Zettlemoyer. Text quality-based pruning for efficient training of language models, 2024. URL https://arxiv.org/abs/2405.01582.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. Slimpajama-dc: Understanding data combinations for llm training, 2024. URL https://arxiv.org/abs/2309.10818.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages, 2024. URL https://arxiv.org/abs/2404.16816.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024. URL https://arxiv.org/abs/2402.00159.

Liping Tang, Nikhil Ranjan, Omkar Pangarkar, Xuezhi Liang, Zhen Wang, Li An, Bhaskar Rao, Linghao Jin, Huijuan Wang, Zhoujun Cheng, et al. Txt360: A top-quality llm pre-training dataset requires the perfect blend, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024a. URL https://arxiv.org/abs/2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil

Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024b. URL https://arxiv.org/abs/2408.00118.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left be-

hind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Yury Tokpanov, Paolo Glorioso, Quentin Anthony, and Beren Millidge. Zyda-2: a 5 trillion token high-quality dataset, 2024a. URL https://arxiv.org/abs/2411.06068.

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. Zyda: A 1.3t dataset for open language modeling, 2024b. URL https://arxiv.org/abs/2406.01981.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. Milu: A multi-task indic language understanding benchmark, 2025. URL https://arxiv.org/abs/2411.02538.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL https://arxiv.org/abs/2411.12372.

Wikimedia Foundation. Wikipedia dumps. https://dumps.wikimedia.org/.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020. URL https://arxiv.org/abs/2010.11934.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. Benchmarking machine translation with cultural awareness, 2024. URL https://arxiv.org/abs/2305.14328.

Xiao Yu, Zexian Zhang, Feifei Niu, Xing Hu, Xin Xia, and John Grundy. What makes a high-quality training dataset for large language models: A practitioners' perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 656–668, 2024.

Xinyan Velocity Yu, Akari Asai, Trina Chatterjee, Junjie Hu, and Eunsol Choi. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources, 2022. URL https://arxiv.org/abs/2211.15649.

Wei Zhang and Alexandre Salle. Native language identification with large language models, 2023. URL https://arxiv.org/abs/2312.07819.

APPENDIX

We are opensourcing the largest high quality Indian Data available. This includes: 2 Trillion tokens of High Quality Indian Language Data spanning 16 Languages, 200 Million Indic Personas (Indic PersonaHub), 300 Million Image Text Pairs (OCR), Taxonomy of 12 Broad Indian Domains, Parallel Long Context Translated Corpora, General SFT Data (Native Indian), Indic MMLU (16 Languages).

A INDIC MMLU

To evaluate state-of-the-art models on Indian languages, we created the Indic-MMLU (Hendrycks et al. (2020)) benchmark, spanning 15 Indian languages and English. Our goal is to examine knowledge and capability transfer from high-resource languages, particularly English, to low-resource Indian languages, as well as to provide a reliable dataset for training and evaluating Indian language LLMs.

Table 7: Absolute Indic MMLU performance (updated with DeepSeek and GPT-OSS models)

Language	DeepSeekR1-0528	DeepSeekV3-0324	DeepSeekV3.1	Gemma-3 27B	gpt-oss-120B High	gpt-oss-120b-med	gpt-oss-120b-low
As	0.6557	0.6638	0.6585	0.5968	0.4585	0.4594	0.4618
Bn	0.7161	0.7293	0.7225	0.6503	0.5190	0.5219	0.5194
En	0.8445	0.8518	0.8569	0.7538	0.7252	0.7255	0.7250
Gu	0.6792	0.6837	0.6795	0.6499	0.4948	0.4967	0.4966
Hi	0.7545	0.7573	0.7533	0.6762	0.5404	0.5400	0.5385
Kn	0.6810	0.6898	0.6847	0.6214	0.4779	0.4803	0.4783
Ml	0.6920	0.7024	0.6958	0.6564	0.5036	0.5020	0.5028
Mr	0.7072	0.7150	0.7120	0.6561	0.5308	0.5295	0.5315
Ne	0.7018	0.7122	0.7021	0.6576	0.5118	0.5145	0.5136
Or	0.6542	0.6639	0.6614	0.6064	0.4662	0.4670	0.4645
Pa	0.6991	0.7072	0.7050	0.6562	0.5091	0.5090	0.5088
Sa	0.6177	0.6252	0.6075	0.5708	0.4026	0.4036	0.4037
Sd	0.4685	0.4822	0.4741	0.4544	0.3462	0.3456	0.3440
Si	0.6077	0.6171	0.6046	0.5625	0.4010	0.4025	0.4030
Ta	0.6911	0.6989	0.6922	0.6447	0.4901	0.4873	0.4900
Te	0.6969	0.7053	0.6954	0.6690	0.5096	0.5100	0.5101
Avg-Indic	**0.6845**	**0.6931**	**0.6869**	**0.6347**	**0.4995**	**0.5000**	**0.4997**

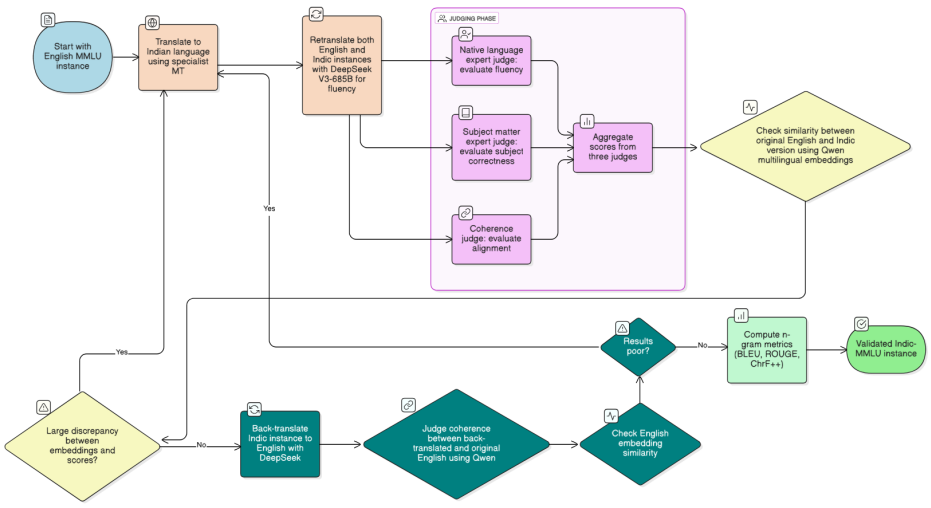

Figure 5: Indic MMLU Creation Workflow

Starting from the original English MMLU test set, we generate high-quality translations into 15 Indian languages. To ensure linguistic naturalness and semantic fidelity, we perform iterative translation and validation using large language models trained for multilingual understanding, as well as expert evaluations considering native language fluency, subject matter accuracy, and cross-lingual coherence. Translations are further validated via back-translation and embedding-based similarity checks to the original English instances, and standard n-gram metrics such as BLEU, ROUGE, and ChrF++ are employed to quantify alignment. The resulting Indic-MMLU benchmark provides a

comprehensive resource for assessing the performance of LLMs on Indian languages and serves as a foundation for downstream data preparation and model evaluation in multilingual settings. The process is also described in Figure 5.

B Data Acquisition

Scaling to millions of books and academic documents across Archive.org, NDLI, and related repositories required not only robust crawling and deduplication pipelines but also a systematic framework for distribution and management of the acquired content. Instead of treating each source as a monolithic corpus, items were organized into orthogonal categories such as language, grade level, provider, and subject domain. This approach ensured that every stage of ingestion preserved provenance and contextual structure, allowing us to both safeguard the integrity of the collections and enable selective sampling for pretraining and supervised fine-tuning.

The NDLI pipeline illustrates this methodology in detail. For school-level content, items were catalogued simultaneously by language, by grade level, and by provider. Table 8 reports the distribution of materials across Indic and English languages, which allowed us to quantify coverage and identify under-represented languages. Table 9 presents the stratification of the same corpus by grade level, making it possible to align training material with the natural curricular sequence from early primary years to higher secondary. Finally, Table 10 organizes the collection by provider, revealing the contributions of individual state boards and institutions. Together, these tables provide complementary perspectives: languages highlight linguistic diversity, grade levels ground pedagogy, and providers reveal provenance. Importantly, these dimensions were treated as orthogonal. Items that were bilingual or cross-listed across grades were preserved with multiple metadata labels and deduplicated only at the item-ID level. This schema-first organization enabled us to design ingestion pipelines that respected the curricular and institutional context of each item, while also facilitating targeted OCR and post-correction workflows adapted to Indic scripts.

Table 8: NDLI School / State-Boards: counts by language (items).

Language	Count
Hindi	4,114
English	3,785
Urdu	1,064
Telugu	755
Sanskrit	657
Kannada	557
Tamil	552
Marathi	481
Gujarati	429
Malayalam	210
Bengali	99
Oriya/Odia	44
Assamese	19
Garo	10
Bodo/Boro	4
Nepali	1
Manipuri	1

Table 9: NDLI School / State-Boards: counts by class/level (items).

Class/Level	Count
Class X	1,926
Class XI	1,611
Class IX	1,533
Class XII	1,483
Class VIII	1,351
Class VII	1,202
Class VI	1,168
Class V	635
Class III	603
Class IV	593
Class I	351
Class II	313

A similar structure was applied to higher-education holdings in NDLI, where the depth and specialization of materials demanded fine-grained partitioning. Table 11 presents the breakdown by content provider, where large-scale curated repositories such as LibreTexts and e-Adhyayan dominate. Table 12 maps the same content to education levels, distinguishing undergraduate, postgraduate, and diploma materials. Table 14 captures subject-specific distributions, surfacing the disciplinary breadth of the collection in areas such as Mathematics, Botany, Chemistry, and Medicine. These independent slices expose different views of the same corpus. Providers reflect provenance, levels capture academic progression, and subjects anchor disciplinary specialization. Maintaining these

Table 10: NDLI School / State-Boards: counts by content provider (items).

Provider	Count
SCERT Telangana	4,247
Raj-eGyan	2,606
Punjab School Education Board	2,465
Gujarat Secondary & Higher Secondary Education Board	788
Jammu & Kashmir State Board of School Education	694
Board of Secondary Education, Madhya Pradesh	520
Karnataka Secondary Education Examination Board	416
NCERT	357
SCERT Kerala	317
SCERT Tripura	100
A. P. Open School Society, Amaravati	85
Assam Higher Secondary Education Council	59
Odisha Primary Education Programme Authority	42
Board of School Education Haryana	33
NCERT — Vocational Education	26
Board of Secondary Education, Odisha	23
Kendriya Vidyalaya ASC Centre(S)	3
Kendriya Vidyalaya Devlali (No. 1)	1

perspectives in parallel allowed us to not only monitor balance and coverage but also to construct evaluation-ready subsets that target either curriculum progression or domain expertise.

Table 11: NDLI Higher Education: counts by content provider (items).

Content Provider	Count
LibreTexts	7,591
e-Adhyayan	6,902
Botanical Survey of India (BSI)	976
Knowledge Unleashed in Multiple Bharatiya Languages (e-KUMBH)	478

Table 12: NDLI Higher Education: counts by education level (items).

Level	Count
Post Graduate	6,901
Under Graduate	1,259
Diploma	146

Table 13: NDLI Higher Education: top subjects (items).

Subject	Count
Mathematics	2,884
Plants (Botany)	900
Commerce, Communications & Transportation	875
Chemistry & Allied Sciences	757
Medicine & Health	755
Engineering & Allied Operations	194
Computer Science, Information & General Works	65
Civil Engineering	59
Other Branches of Engineering	59
Plants noted for characteristics & flowers	45
Others	208

This methodology of multi-dimensional cataloging extended across all sources of acquisition, including Archive.org and government portals. At each step, metadata was normalized into consistent schemas and tabular reporting was applied, producing structured distributions analogous to the NDLI examples shown here. This ensured that every dataset entering the pipeline was not merely collected but carefully managed, interpretable, and auditable. By grounding acquisition in struc-

Table 14: NDLI Higher Education: top subjects (items).

Subject	Count
Mathematics	2,884
Plants (Botany)	900
Commerce, Communications & Transportation	875
Chemistry & Allied Sciences	757
Medicine & Health	755
Engineering & Allied Operations	194
Computer Science, Information & General Works	65
Civil Engineering	59
Other Branches of Engineering	59
Plants noted for characteristics & flowers	45
Others	208

tured distribution, we transformed raw web-scale scraping into a governed resource that could be harnessed for curriculum-aware pretraining, subject-specific fine-tuning, and rigorous multilingual evaluation.

C Data Curation

High-quality and linguistically diverse data is foundational for robust multilingual language models. In addition to the summary provided in the main paper, we provide a comprehensive view of our curation process, its practical implementation, and the resulting improvements in dataset quality and safety.

C.1 Pipeline Overview

The complete data curation workflow is presented in Figure 6. For clarity, this figure provides a visual reference of the sequential processing steps, including quality scoring, filtering, language identification, deduplication, toxic content removal, and PII redaction. Readers are encouraged to refer to this diagram to understand the structural flow of data through the pipeline.

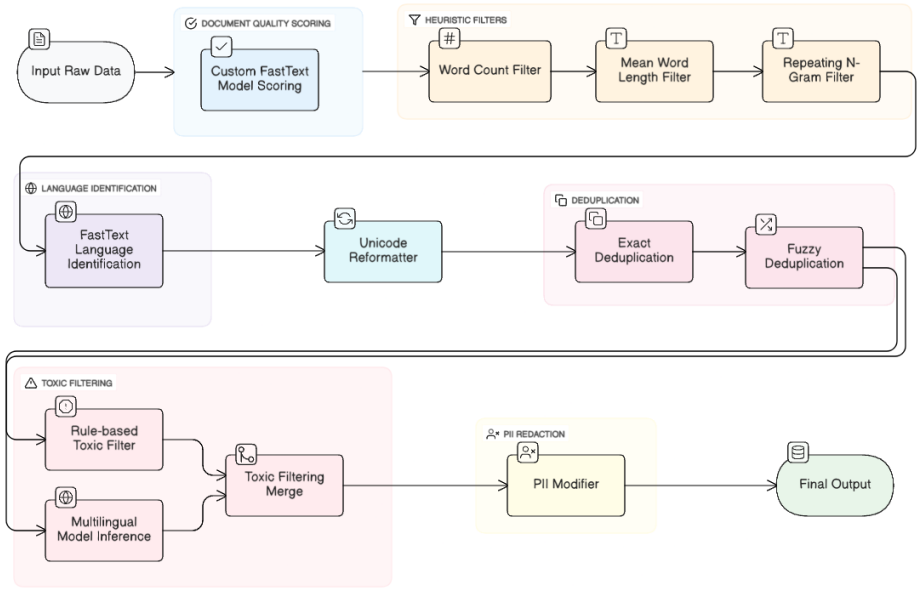

Figure 6: Curation Pipeline

C.2 CURATED EXAMPLE: ASSAMESE TEXT

To illustrate the practical impact of the pipeline, we present an Assamese text extracted from Common Crawl, showing the before-and-after effects of curation (Figure 7). The raw text contains inline HTML, formatting artifacts, repeated symbols, and inconsistent spacing. After curation, these artifacts are removed, resulting in a clean, readable, and linguistically faithful version. The curated text preserves all semantic and contextual information, demonstrating how our approach enhances text quality for low-resource Indic languages.

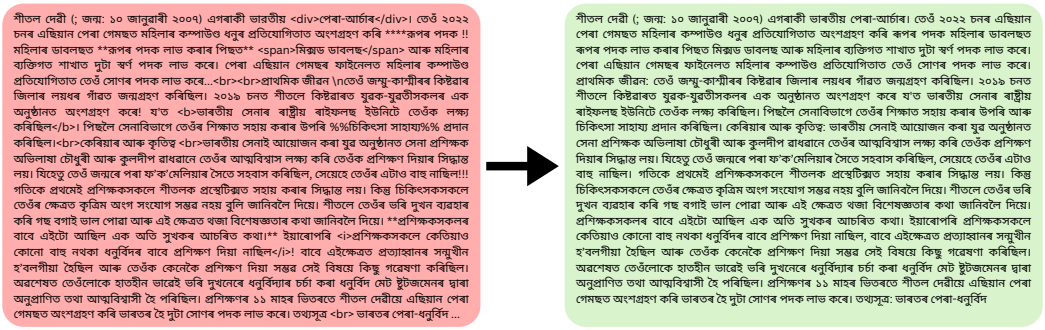

Figure 7: Conventional vs Curated Comparison

C.3 TRAINING SETUP OF PARAM-1

The benchmark results in Table 2 illustrate the clear advantage of curated training data over conventional data, with the PARAM-1 2.9B model achieving substantial improvements across multiple downstream tasks. Importantly, the only difference between the conventional and curated experiments is the data; all other training conditions remain identical. The PARAM-1 model is a causal language model trained initially on English and Hindi data (architecture details are provided in Table 15). This setup allows the model to effectively leverage curated data, producing substantial improvements while maintaining identical hyperparameters and training regime.

Architecture attributes	Values
Model Architecture	causal-language-model
Hidden size	2048
Intermediate size	7168
Max Position Embeddings	2048
Num of Attention Heads	16
Rope theta	10000
Num of Hidden Layers	32
Num of Key Value Heads	8
Activation Function	fast-swiglu
Attention Type	Grouped-query attention
Precision	bf16-mixed

Table 15: Architecture Details of PARAM-1

C.4 TOXIC FILTERING EVALUATION

Beyond conventional quality improvements, curated training data significantly enhances model safety. PARAM-1, trained on the curated dataset, demonstrates superior performance in minimizing harmful content generation, particularly in sensitive linguistic and cultural contexts. To quantify this, we evaluated toxicity using the Toxigen benchmark via LLM360's Safety360 suite, which provides both explicit and subtle adversarial prompts across identity-based and general categories. The curated PARAM-1 model achieves lower toxicity rates than comparable multilingual baselines, in-

cluding SARVAM-1, LLaMA3.2-9T-3B, and Gemma2-2T-2B, demonstrating that improvements in data curation directly translate into safer language model outputs.

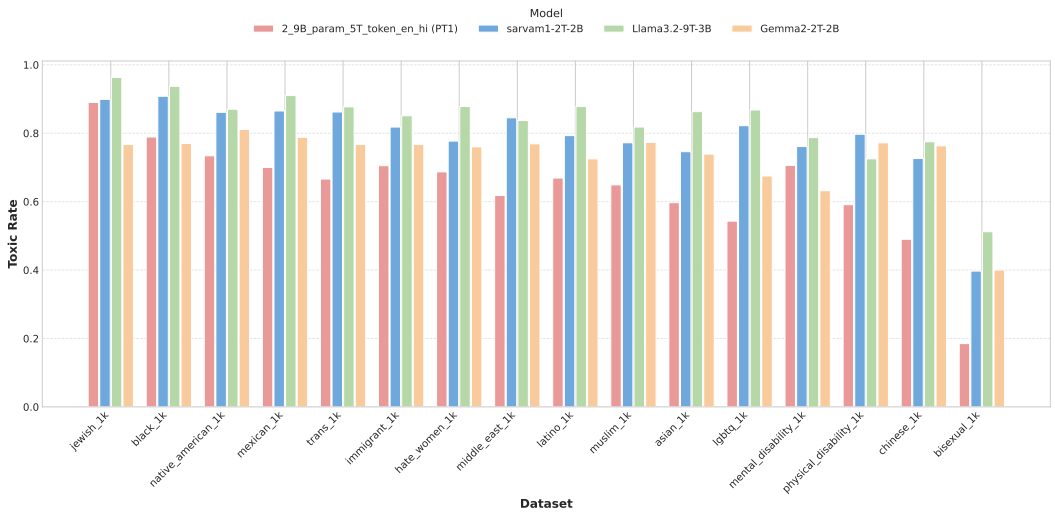

Figure 8: Toxicity Comparison

The evaluation procedure involves generating model outputs with curated Toxigen templates, followed by classification using a RoBERTa-based detector fine-tuned for nuanced and context-dependent toxic language. Across 16 datasets, encompassing both neutral and adversarial examples, PARAM-1 consistently maintains lower toxicity rates while avoiding stereotype amplification and identity-based abuse. These results, visualized in Figure 8, highlight that the curated training process fosters models that are not only more accurate but also culturally aware and safe, capable of generating neutral or helpful responses even in challenging scenarios.

Together, these results underscore the dual benefit of our data curation pipeline: it improves both task-specific performance, as evidenced in Table 2, and the ethical safety of model outputs, establishing a strong foundation for responsible deployment in multilingual and low-resource contexts.

D OCR PIPELINE

D.1 LLM-ASSISTED QUALITY EVALUATION

To rigorously assess the semantic fidelity of OCR outputs beyond conventional metrics, we designed an evaluation framework where large language models serve as quality judges. While word error rate (WER) and character error rate (CER) provide useful signals for raw OCR accuracy, they are insufficient when text undergoes quality enhancement, as sentence reordering, ligature repair, and other postprocessing interventions may reduce traditional scores while preserving or improving semantic content. To address this, we employ state-of-the-art LLMs such as GPT-OSS-120B, Deepseek, and Qwen, to perform a multi-stage evaluation of both OCRed and quality-enhanced outputs.

Our experiments focus on Indian language benchmarks within the ISOB-Small dataset. Each LLM is prompted to compare the original ground truth text in English with the corresponding OCRed and postprocessed outputs in their native scripts. The models provide consistency scores reflecting the degree to which the enhanced text preserves the meaning of the ground truth. In parallel, embedding similarity metrics are computed using multilingual embeddings, capturing semantic alignment between the original and processed text in the native language space. To evaluate cross-lingual fidelity, we translate both ground truth and enhanced text into English using Deepseek and compare them using Qwen, complemented by embedding similarity scores in English. Finally, translation-level evaluation metrics such as BLEU, ROUGE, and CHRF++ are incorporated to provide a holistic view of text quality, and image-level embedding similarity is employed to measure structural fidelity between pages reconstructed from raw ground truth versus OCRed and enhanced outputs.

D.2 Multi-Stage OCR Processing and Validation Pipeline with Instruction-Grade Prompts

Input: Digitized or undigitized document images (low-quality, noisy, or artifact-rich)
Output: Validated OCR text with associated confidence and quality labels

Step 1: Pre-Processing

1.1 Error Identification using VLMs

Model: Lightweight VLM (e.g., Qwen-VL-7B)
Task: Detect artifacts, orientation, and readability issues in scanned pages.
Prompt:

```
You are a document analysis system.
Given this scanned page, identify the following issues:
1. Page orientation (normal, rotated, upside-down)
2. Presence of noise, blur, or watermarks
3. Regions of non-text (stamps, illustrations, smudges)
4. Overall readability score (High, Medium, Low)
Return the issues as a structured JSON object.
```

1.2 Artifact Removal & Enhancement

- Apply preprocessing checks for orientation, noise, blur
- Apply Super-Resolution (SRGAN) for low-resolution pages
- For severe degradation, use Qwen Image Edit with targeted prompts

Prompt (Qwen Image Edit):

```
You are an image enhancement system.
Task: Improve readability of this scanned document.
Instructions:
- Sharpen text edges
- Remove background noise and smudges
- Correct orientation if tilted
- Increase resolution while preserving textual structure
Return the enhanced page without altering the content.
```

Step 2: OCR Generation with Human-in-the-Loop

- Use an ensemble of specialist OCR models (per script/language) + generalist VLMs for layout interpretation
- Periodic human calibration on sampled pages ensures adaptation per language/domain

Prompt (Generalist VLM for layout):

```
You are an OCR layout assistant.
Given this scanned page, output:
1. The document layout (columns, tables, figures)
2. Logical reading order of text blocks
3. Any script/language hints detected
Return in structured JSON to assist OCR alignment.
```

Step 3: Post-OCR Quality Enhancement

- Apply rule-based filters (dictionary constraints, script normalization)
- Use LLM-based post-correction for grammar, consistency, and semantic alignment

Prompt (LLM Post-Correction):

STEP 4: VALIDATION & CONSISTENCY CHECKING

4.1 RECONSTRUCTION WITH HOCR + STYLE TRANSFER

- Reconstruct page from hOCR using Qwen Image Edit

Prompt (Qwen Image Edit - Style Transfer):

```
You are an image reconstruction system.
Task: Using the provided hOCR, generate an image that visually resembles the original
Constraints:
- Preserve layout, fonts, and formatting
- Apply natural degradation styles common in manuscripts (faded ink, paper texture)
Return the synthetic page.
```

4.2 EMBEDDING SIMILARITY CHECK

Compute cosine similarity between embeddings of original vs reconstructed images.

4.3 REASONING-BASED VALIDATION WITH VLMS

- Input both original and reconstructed pages to reasoning-capable VLMs (e.g., Qwen-32B–235B)
- Generate independent interpretive trajectories of page content

Prompt (Reasoning VLM):

```
You are a reasoning OCR evaluator.
Given this scanned page, describe step by step:
1. What the page contains (title, paragraphs, tables, etc.)
2. The key semantic content (entities, topics, structure)
Return a reasoning trajectory of what you understand from this page.
```

4.4 TRAJECTORY COMPARISON WITH LLMS

- Provide reasoning trajectories (original vs reconstructed) to a strong LLM for similarity scoring

Prompt (Text LLM Comparison):

```
You are a similarity evaluator.
Input: Two reasoning trajectories of the same page (original vs reconstructed)
Task:
1. Compare their semantic overlap
2. Assign a similarity score (0{100)
3. Provide justification for the score
4. Bucket the result into {Good, Acceptable, Bad}
Return output as JSON: {score, justification, bucket}.
```

4.5 SCORE-REASON CONSISTENCY CHECK

- Use a secondary LLM to verify if justification aligns with score

Prompt (Consistency Check LLM):

```
You are a validation assistant.
Input: {score, justification} from similarity evaluator
```

```
Task: Verify if the justification logically supports the score
Return either "Consistent" or "Inconsistent".
```

STEP 5: HUMAN EXPERT REVIEW AND MANUAL POST CORRECTION

Pages flagged as low similarity or inconsistent are routed to human linguists for manual validation.

FINAL RESULT

An orchestrated OCR pipeline integrating artifact detection, super-resolution, ensemble OCR, post-correction, reconstruction-based validation, reasoning-VLM alignment, and human-in-the-loop oversight. All model interactions are guided by instruction-grade prompts f

D.3 ISOB: INDIC SYNTHETIC OCR BENCHMARK

Further, to evaluate OCR performance on Indic scripts, we developed the Indic Synthetic OCR Benchmark (ISOB-Small) as a controlled proxy for real-world challenges. Directly using copyrighted print materials poses legal and ethical constraints, and many low-resource Indic languages remain underrepresented in publicly available corpora. To address this, ISOB-Small is synthetically generated to emulate common complexities observed in scanned documents, including multi-column layouts, tables, figures, mathematical expressions, watermarks, folds, font variations, and degradations such as blur, shadows, and uneven lighting. By generating these diverse artifacts, the benchmark reproduces the types of OCR errors that occur in real-world digitized Indic texts, while remaining fully copyright compliant.

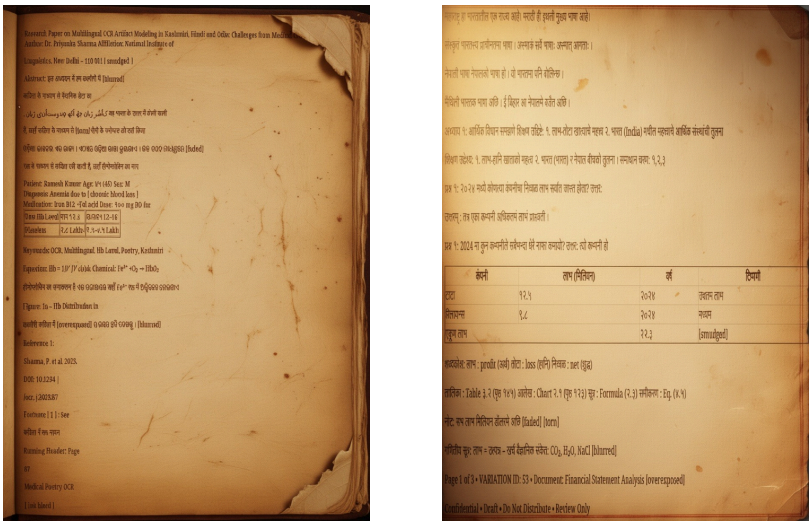

Figure 9: Samples from ISOB Benchmark

Since a significant portion of our corpus originates from copyrighted materials obtained under formal MoUs and partnerships, we are unable to release those pages directly as benchmarks. However, recognizing the complexity and challenges inherent in such offline digitized documents—and to support the research community in addressing them—we are introducing the first version of the **Indian Synthetic OCR Benchmark – Small-Hard**.

Future releases will expand to include:

- **Indic-Real-OCR Benchmarks:** Licensed for public release, categorized into Easy, Medium, and Hard difficulty levels, and available in Small, Medium, and Large sizes.
- **Indic-Synthetic-OCR Benchmarks:** Similarly structured across Easy, Medium, Hard levels and Small, Medium, Large dataset sizes.

Input: Seed corpus of OCR'd pages in hOCR format

Output: Synthetic benchmark dataset of hard-to-OCR images with ground truth hOCR and language tags

STEPS FOR DATASET CREATION

Step 0. Initialize Seed Corpus
Use existing OCR'ed pages (hOCR format) as the starting corpus.

Step 1. Hard Page Identification
Filter pages using OCR confidence scores:

- Discard pages with too low confidence (often whitespace, empty, or very low-text pages)
- Use Qwen-VL-7B grounded with the page hOCR to predict difficulty
- Select pages predicted as hard-to-OCR → `hOCR_1` (hard page set)

Step 2. Language Selection
Randomly choose a set of 3–10 languages from a pool of 22 Indian languages.

Step 3. Artifact Taxonomy Extraction
Build a taxonomy (list) of "hard artifacts" from the seed hOCR corpus using LLMs, grounded in the reality of complex documents. Examples include: multi-column layouts, dense tables, handwriting inserts, overlapping scripts, reading orders, equations, figures, pie charts, complex tables, etc.

Step 4. Synthetic hOCR Augmentation
Augment each selected `hOCR_1` using:

- Chosen languages (Step 2)
- Artifact templates (Step 3)

Produce new enriched hOCR documents → `hOCR_2`, which serves as ground truth.

Step 5. hOCR to Visual Conversion
Render each `hOCR_2` into PDF/image format.

Step 6. Style Transformation (Prompt Pool + Image Editing)
Construct a prompt pool describing Indian manuscript styles, books, literature, and other domain-specific styles. For each page:

- Sample a prompt from the pool
- Apply Qwen image editing with the prompt to transform the visual style

Step 7. Image Processing Augmentation
Apply low-level transformations to further increase difficulty, including: orientation changes, contrast/brightness shifts, noise, blur, distortions, etc.

Step 8. Storage & Annotation
Save final images with metadata:

- Associated ground truth hOCR (`hOCR_2`)
- Language tags (Step 2)
- Style/augmentation metadata

End Result: A structured, multilingual, style-rich, artifact-heavy synthetic OCR benchmark that systematically captures hard-to-recognize text cases.

ISOB-Small spans 22 Indian languages and consists of 110 pages designed to stress-test OCR systems, particularly their ability to handle script-specific features such as ligatures, conjunct consonants, and diacritics. Figure illustrates the ISOB creation pipeline, detailing the process from synthetic layout generation to language-specific text rendering and controlled degradation application. A representative example page from ISOB-Small is shown in Figure 9.

In addition to providing the benchmark, we are releasing both the dataset and the generative recipes used to construct it. This allows researchers to extend ISOB-Small or create proxy benchmarks for other low-resource scripts, facilitating reproducible and legally compliant experimentation in Indic OCR research. By offering an open-source reference framework, ISOB-Small provides a foundation for future work on model evaluation, preprocessing strategies, and domain adaptation.

E Translation Pipeline

E.1 Chunking Strategy for Long Contexts

Translation of long and complex texts, particularly in domains such as mathematics, formal proofs, and code, presents unique challenges due to context dependencies and sequence length limitations of large language models. To address this, we implement a hierarchical chunking strategy. Texts are divided into contiguous segments based on token counts and logical units, ensuring that each chunk remains within the model's context window. To preserve semantic dependencies across chunks, each segment is provided with summaries of preceding segments, enabling the model to maintain continuity of reasoning. For highly structured content such as multi-step proofs or code blocks, overlapping tokens between consecutive chunks are included to prevent loss of critical context. This chunking strategy ensures that long documents are translated coherently, with minimal information loss and faithful reproduction of logical and syntactic structures.

Following initial translation, outputs undergo an LLM-based post-correction phase. We select only models that demonstrate strong performance on Indic MMLU benchmarks (see Table 5) to ensure that postprocessing leverages linguistic proficiency in low-resource Indic languages. The post-correction phase is designed to repair grammatical inconsistencies, improve sentence flow, and enforce naturalness while strictly preserving the original meaning. LLMs are prompted with language-specific instructions emphasizing grammatical accuracy, syntactic correctness, and culturally appropriate usage, without introducing English or Hindi mixing. This targeted refinement significantly reduces residual errors that arise from direct machine translation, particularly in complex or idiomatic constructs.

E.2 Example Prompts and Postcorrected Output

Below we present the prompt template used for LLM-based post-correction, along with processed output. The prompts are designed to guide the model toward high-fidelity, grammatically correct, and natural-sounding translations. A sample of text before and after postcorrection is shown in 10.

Prompt Template

Role and Context: You are an expert linguist specializing in the {language} language with deep understanding of grammar, syntax, and natural language use.
Task: Transform {language} text that is poorly structured, grammatically incorrect, awkwardly translated, or unnatural into well-formed, grammatically correct, and natural-sounding {language} text.
Input Text: '{input_text}'
Output Requirements:

- Return complete rephrased text with no omissions wherever needed
- Never return empty responses
- Maintain original language with no English/Hindi mixing
- Focus on grammatical correctness and natural flow
- Do not provide explanations, notes, or meta-commentary
- Keep the length close to the original text

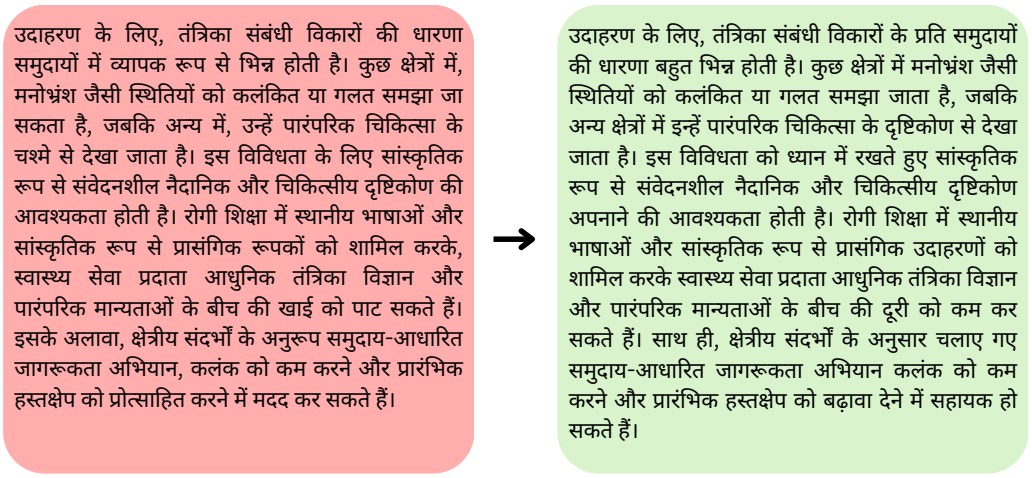

Figure 10: Toxicity Comparison

E.3 Multilingual Translation and Post-Correction Pipeline for Indic Languages

Input: Multilingual datasets (general domain, technical knowledge, low-resource Indic languages, complex modalities)
Output: High-quality Indic translations with human-validated corrections and long-context reasoning

Step 1: Data Augmentation and Diversification

Augment Indian language datasets with:

- Parallel data from other languages
- Complex modalities (code, math, STEM, proofs)
- Diverse domain-specific tasks

Rationale: Improves cross-domain generalization and strengthens model reasoning abilities.

Step 2: Initial Translation (Ensemble Generation)

Generate translations using ensembled models:

- Specialist models (domain/language-tuned) for technical fidelity
- Generalist LLMs for broad fluency and adaptability

Observation: Feeding specialist model outputs to generalist models generally improves translation quality (human judgment analysis).
Output: Multiple candidate translations per segment.

Prompt [Generalist Translation]:

Step 3: Post-Correction and Quality Enhancement

3.1 Semantic Consensus via LLM-as-Judge

Use strong multilingual LLMs (e.g., Qwen-235B, DeepSeek, GPT-OSS) to rank and score translations based on:

- Semantic fidelity
- Grammatical correctness
- Fluency
- Domain preservation

Prompt [Semantic Judge]:

```
You are a multilingual semantic evaluator.
Input: Source text + multiple candidate translations.
Task: Rank translations for:
(i) semantic fidelity
(ii) grammar/fluency
(iii) technical correctness
Return JSON {best_translation, justification}.
```

3.2 Back-Translation for Robustness

- Translate Indic output back into English
- Compare embeddings with original English source
- Use LLM as judge to finalize similarity assessment

Step 4: Human Evaluation and Finalization

- Low-score cases are sent for human verification and post-correction
- Human evaluators validate translations chosen via consensus
- Calibration ensures domain-sensitive correctness in Indic contexts (legal, STEM, literary)

Step 5: Long-Context Chunking Strategy

For documents with long sequences (20K–25K tokens):

- Apply chunking strategy for autoregressive generation
- Maintain coherence across segments by overlapping context windows

Prompt [Chunked Translation]:

```
You are a long-context translator.
Input: Segment of a long document (with overlapping context from previous segment)
Task: Translate into {Indic language}, ensuring continuity with prior segments
Return translation only.
```

Rationale: Handling complex technical/multilingual data improves reasoning abilities and long-context coherence **??**.

Step 6: Consolidation and Verification

- Merge chunked outputs into final translated document
- Validate coherence using LLM reasoning + embedding similarity across boundaries

Prompt [Coherence Validator]:

```
You are a coherence evaluator.
Input: Consecutive translated chunks
Task: Verify continuity of meaning, terminology consistency, and semantic flow
Return verdict: {Consistent, Inconsistent, Requires Edit}.
```

Step 7: Expert Review and Benchmarking

- Flag low-confidence cases for human linguist review
- Benchmark translations against:
  - Semantic LLM scores
  - Embedding similarity metrics
  - Human evaluation reports

Final Result

A scalable translation pipeline that combines augmentation, ensemble generation, post-correction with state-of-the-art LLMs, embedding-based validation, back-translation, and chunking strategies to deliver high-quality Indic translations across general and technical domains. Human-in-the-loop evaluation ensures reliability and domain fidelity.

F Synthetic Augmentation, Rewriting and Data Distillation

F.1 Data Distillation Pipeline

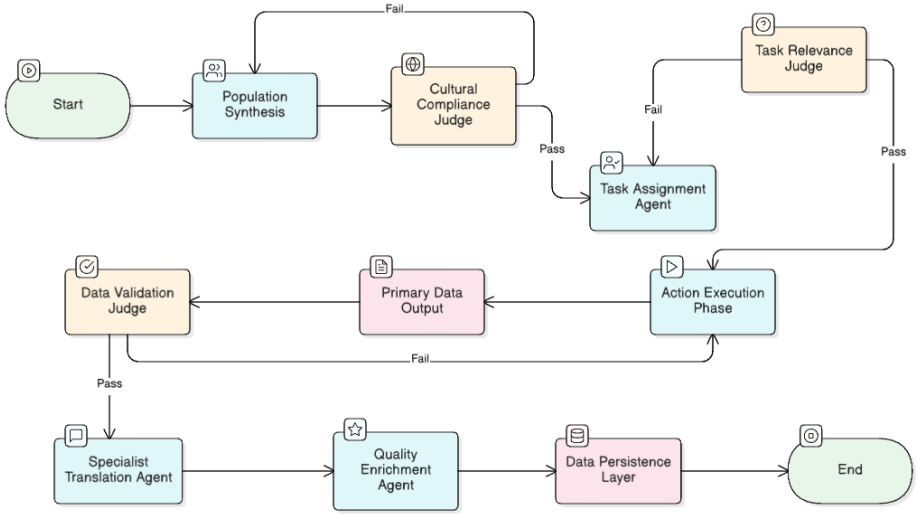

Figure 11: Data Distillation Pipeline

The data distillation pipeline, illustrated in Figure 11, begins with a large-scale population synthesis phase in which over 370 million personas are generated across more than 1400 broad domains. These raw personas capture a wide breadth of potential identities and expertise but remain unfiltered. To ensure cultural and societal alignment, each persona is first reviewed by a cultural compliance judge. Personas failing this evaluation are recycled for regeneration, while those passing proceed to the task assignment agent, which pairs each persona with relevant domains, contexts, and functions. The assigned persona-task pairs are then evaluated by a task relevance judge, ensuring that the tasks are logically coherent and meaningful. If mismatches are detected, the persona-task pair is sent back for reassignment, maintaining rigorous quality control throughout the pipeline.

Once a persona and its task are approved, the pipeline moves to the action execution phase, generating primary outputs that serve as the first substantive data artifacts. These outputs undergo scrutiny by a data validation judge, which checks for factual accuracy, coherence, and overall alignment. Data passing validation is forwarded to a specialist translation agent to produce high-quality outputs in Indic languages or to refine stylistic and linguistic properties. A quality enrichment agent further enhances fluency, context, and naturalness before the outputs are stored in the data persistence layer.

This layered structure, with iterative feedback loops at each stage, ensures that only thoroughly vetted and enriched material becomes part of the training corpus, providing a strong foundation for the subsequent Indianization of personas process.

An essential dimension of our distillation process is the Indianization of personas, which ensures that synthetic datasets not only maintain linguistic precision but also embed culturally grounded knowledge, values, and perspectives. Without this, even the most linguistically accurate synthetic corpora risk reproducing Western biases and overlooking the cultural richness of Indian society. By systematically transforming generic or Western-centric personas into contextually Indian personas, we ensure that the synthetic text aligns with national identity, societal norms, and domain-specific expertise relevant to Indian readers.

To achieve this, we introduce a three-stage persona Indianization pipeline, where each stage incrementally deepens the cultural embedding of the persona while retaining the intellectual and professional integrity of the original role.

STAGE 1: INDIANIZED PERSONA GENERATION

The first stage modifies existing personas to explicitly reflect Indian cultural values and national alignment. The prompt carefully preserves the core traits of the original persona while ensuring they embody patriotism and avoid anti-Indian narratives. This guarantees both authenticity and contextual appropriateness.

Stage 1 Prompt

Modify the following persona to make it culturally Indian while preserving their core personality traits. Ensure the persona remains aligned with Indian nationalism, avoids any anti-Indian government stance, and embodies patriotism towards India.
{original_persona}
Provide the generated response (strictly in English language) in the following key, value format without any header.
Format:
ps: value of Indianized persona — gd_1: Indianized value of general domain (1%) — sp_1: Indianized value of specific domain (1%) — gd_01: Indianized value of general domain (0.1%) — sp_01: Indianized value of specific domain (0.1%)

Sample excerpt: A scientist working on embryonic stem cell research is transformed into a dedicated Indian scientist whose ethical advocacy is framed through contributions to India's healthcare ecosystem, with domain anchors adjusted to reflect Indian policy, ethics, and biomedical priorities.

STAGE 2: THOUGHT-PROVOKING QUESTION GENERATION

The second stage builds upon the Indianized persona by prompting the model to produce **deeply reflective, domain-specific questions**. These questions are tailored to stimulate reasoning in the Indian context, ensuring that generated text moves beyond rote repetition and embraces the nuances of Indian society, ethics, and challenges.

Stage 2 Prompt

Craft a single, thought-provoking and mind-triggering question in {domain} that inspires deep reflection and invites a broad exploration of ideas, perspectives, and reflections, particularly within the Indian context. Provide only the question (strictly in English language), without any additional commentary or explanation.

Sample excerpt: In the case of Stem Cell Ethics, the generated question becomes: *How should India navigate the ethical complexities of stem cell research and therapy, balancing the potential for groundbreaking medical advancements with the moral considerations surrounding the use of embryonic stem cells, informed consent, and equitable access to treatments?*

STAGE 3: FINAL DATA GENERATION

The final stage transforms the enriched persona and question into a large-scale text generation task, producing high-quality passages of over 900 words. These outputs are not arbitrary; they are designed to read naturally, flow logically, and resonate with Indian audiences while maintaining global scientific relevance. The prompts explicitly instruct the model to maintain coherence, persuasive yet neutral tone, and Indian cultural grounding.

Stage 3 Prompt

You are {remove_markdown(output.persona)} Your role is to engage with users based on your expertise. Stay within your domain and maintain the persona's tone and expertise.
CONTEXT
The need for this dataset stems from the desire to uphold a standard of excellence in English language content within the {remove_markdown(output.domain)} field. By compiling a diverse range of well-structured and authentic texts, this collection will help maintain a rich linguistic resource that supports clarity, readability, and contextual accuracy in various forms of communication.
OBJECTIVE
I want you to generate text paragraphs strictly in English language with 900+ words for {remove_markdown(output.domain)} that is easy to read, flows naturally, and sounds like it was written by a human. Generated text data should mimic real world data so that it can also be used to improve research and innovation. Use clear transitions between sentences and paragraphs while maintaining a consistent narrative or argument ensuring a logical progression of thought. Ensure the writing is engaging and not mechanically repetitive. Question: {remove_markdown(output.generated_text)}
STYLE
Follow the simple writing style common in communications. Be persuasive yet maintain a neutral tone. Avoid sounding too much like a sales or marketing pitch.
AUDIENCE
The primary audience is of Indian origin, so content should incorporate cultural familiarity, societal norms, and linguistic nuances relevant to Indian readers.
RESPONSE
Generate a well-structured and engaging piece of content adhering to the above parameters. The writing should feel natural, contextually appropriate, and resonate with the target audience.

Sample excerpt: The persona of the Indian stem cell scientist produces a comprehensive essay that seamlessly blends scientific reasoning with India's healthcare priorities, ethical frameworks, and aspirations of global leadership, creating content that is linguistically robust and culturally authentic.

This three-stage design ensures that synthetic data generation moves beyond surface-level translation. Instead, it becomes a **culturally adaptive process** that embeds Indian perspectives into the very core of persona-driven content creation. By coupling Indianized personas with reflective questions and extended content generation, the pipeline produces training material that is both globally competitive and deeply rooted in local relevance.

F.2 EXPANDED QA EXTRACTION PIPELINE

The QA extraction pipeline operationalizes the transformation of unstructured Indic text into high-quality instruction data through four stages.

Context-Aware Chunking. Raw text is segmented into meaningful spans, typically between 1000 and 4000 tokens, with rules preventing mid-sentence breaks and preserving logical coherence. This ensures that each segment is interpretable as a standalone unit.

Relevance Checking and Domain Classification. Each chunk is validated for cultural or societal relevance to the Indian context, filtering out ephemeral or narrowly technical material. Valid segments are then assigned to domains such as Healthcare, Finance, History, Culture, BFSI, Educa-

tion, Governance, Law, News, Sports, and Tourism. In practice, 1121 chunks were sourced from wikipedia_indic, 619 from dharmawiki, and 4775 from other domains combined.

Self-Contained Question Generation. From validated chunks, fully independent questions are produced. These include general explanation questions, commonsense reasoning questions, causal reasoning questions, and open-ended prompts. Each question is constructed to stand alone without requiring reference back to the original text.

Multi-Fidelity Answer Generation. Answers are created in two complementary forms. Crisp answers provide a concise response, while detailed answers, typically three to five sentences, supply explanatory context and elaboration.

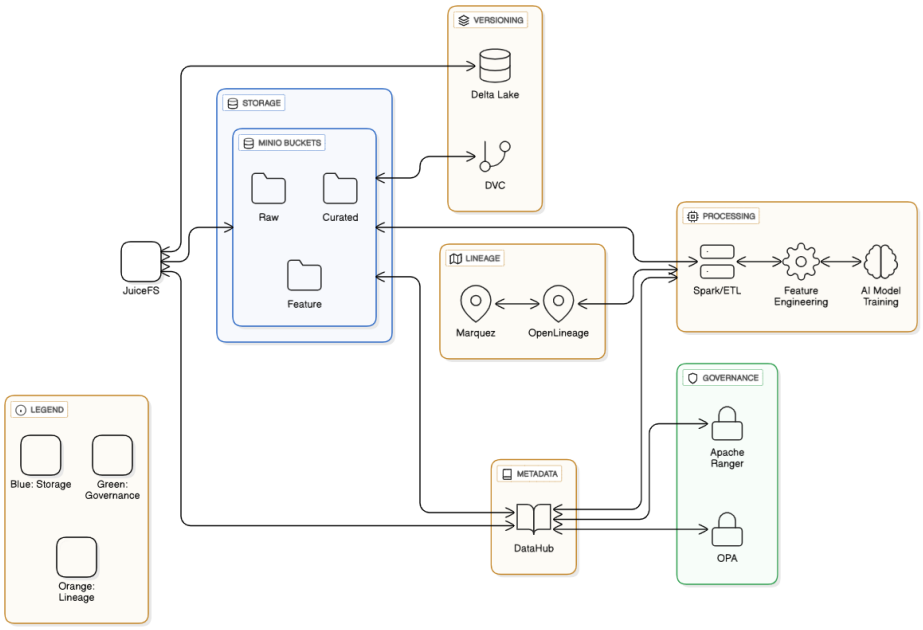

Figure 12: Data Governance Pipeline

By combining synthetic article generation with this structured QA extraction pipeline, the dataset achieves both breadth and depth. The result is a resource that is factually grounded, instruction-ready, and culturally resonant, making it uniquely suited for fine-tuning models in the Indian context.

Task	Score Name	Open Source SFT	In-house SFT Data Recipe
hellaswag	acc_norm, none	70.47	73.07
hellaswag_hi	acc_norm, none	44.01	44.59
global_mmlu_full_en	acc, none	37.89	37.4
global_mmlu_full_hi	acc, none	31.43	31.65
mmlu_pro	exact_match, custom-extract	5.23	8.73
piqa	acc_norm, none	78.24	79.22
winogrande	acc, none	62.04	62.19
truthfulqa_gen	bleu_acc, none	35.74	37.7
truthfulqa_mc1	acc, none	27.17	29.74
cb	acc, none	30.36	57.14
milu_English	acc, none	35.95	37.19
milu_Hindi	acc, none	28.87	32.26
sanskriti_states	acc, none	55.13	55.91

Table 16: Comparison of Open Source SFT and In-house SFT Data Recipe across different tasks.

G DATA ORGANIZATION

Our lakehouse framework emphasizes structured, auditable, and taxonomy-driven data organization for Indic corpora. Every asset is annotated with domain, language and script, modality, license, source, sensitivity, quality tier, stage, and lineage metadata. Domains include Agriculture, Culture, Education, News, Business, Healthcare, Sports, Law, Governance, Tourism, and BFSI. Languages and scripts include Devanagari, Latin, Malayalam, Telugu, Hindi, Marathi, Malayalam, and English. Modality spans text, PDF, images, audio, and code. Quality tiers encode OCR and readability, while license and sensitivity tags ensure safe promotion for downstream use. All metadata is captured in DataHub, enabling auto-generated dataset cards that maintain descriptive, operational, and governance-aligned documentation.

The data governance pipeline, illustrated in Figure 12, operationalizes these annotations. Raw data enters MinIO-backed zones (Raw, Curated, Feature) via JuiceFS and is processed through Spark ETL pipelines and feature engineering modules before being ingested into AI training pipelines. DataHub captures full metadata and lineage, while governance policies enforced through Apache Ranger and OPA evaluate license safety, PII constraints, and domain-specific rules at every promotion or sampling stage. OpenLineage and Marquez track per-step events, and Delta Lake with DVC provide versioned, reproducible snapshots. This structure ensures each asset's provenance, quality, and compliance are fully traceable. Assets flow from ingestion through curated and feature layers under strict policy and lineage supervision, forming a reliable foundation for persona-driven data distillation.

H TABLES

Figure 13: IN22 DeepSeek V3.1 Think Results

Language	chrF	chrF++	BLEU
asm_Beng	24.005	22.643	5.265
ben_Beng	32.066	29.752	6.429
guj_Gujr	27.137	27.827	8.883
hin_Deva	48.308	46.546	21.875
kan_Knda	32.011	29.178	5.046
mai_Deva	29.344	26.375	5.106
mal_Mlym	22.474	20.668	2.588
mar_Deva	41.255	37.508	10.064
npi_Deva	40.244	36.064	8.542
ory_Orya	8.956	10.671	1.635
pan_Guru	29.568	29.171	10.314
san_Deva	28.630	24.424	3.188
tam_Taml	38.307	33.895	5.342
tel_Telu	29.887	27.682	4.587
urd_Arab	42.839	41.726	20.242

Figure 14: Qwen-235B-Thinking Results

Language	chrF	chrF++	BLEU
asm_Beng	44.429	40.073	11.380
ben_Beng	50.108	45.314	13.978
guj_Gujr	49.926	46.752	17.691
hin_Deva	55.125	53.073	27.112
kan_Knda	50.568	45.054	11.002
mai_Deva	43.797	39.228	9.478
mal_Mlym	48.123	41.771	8.109
mar_Deva	50.375	46.008	15.681
npi_Deva	49.657	44.626	12.875
ory_Orya	46.758	42.051	11.833
pan_Guru	46.099	43.317	17.685
san_Deva	33.216	28.015	4.515
tam_Taml	53.980	47.581	11.806
tel_Telu	51.580	46.693	13.640
urd_Arab	57.824	55.589	31.071

Figure 15: IN22 Second Column Results

Language	chrF	chrF++	BLEU
asm_Beng	39.801	35.655	8.061
ben_Beng	46.097	41.515	11.770
guj_Gujr	44.736	41.539	13.008
hin_Deva	51.896	49.802	23.271
kan_Knda	44.180	38.938	7.252
mai_Deva	36.728	31.566	5.810
mal_Mlym	43.082	37.184	6.308
mar_Deva	44.090	40.026	10.779
npi_Deva	46.570	41.610	10.972
ory_Orya	41.326	36.657	7.473
pan_Guru	41.706	39.075	14.026
san_Deva	30.653	25.737	3.103
tam_Taml	49.523	43.561	10.066
tel_Telu	47.718	42.956	11.615
urd_Arab	51.807	49.455	23.868

Figure 16: Flores DeepSeek V3.1 Think Results

Language	chrF	chrF++	BLEU
asm_Beng	42.407	39.104	9.204
ben_Beng	48.999	44.298	13.148
guj_Gujr	51.236	47.953	19.540
hin_Deva	55.346	53.115	28.656
kan_Knda	52.856	47.556	13.582
mai_Deva	43.479	39.588	11.010
mal_Mlym	52.733	46.863	12.067
mar_Deva	49.050	44.797	14.239
npi_Deva	52.144	47.234	14.086
ory_Orya	49.054	44.340	12.638
pan_Guru	48.814	46.801	21.626
san_Deva	31.091	25.373	1.190
tam_Taml	55.253	48.895	12.750
tel_Telu	53.380	48.678	15.974
urd_Arab	48.353	46.047	20.454