MILA (Multilingual Indic Language Archive): A Dataset for Equitable Multilingual LLMs

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 pretraining 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.[a]

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

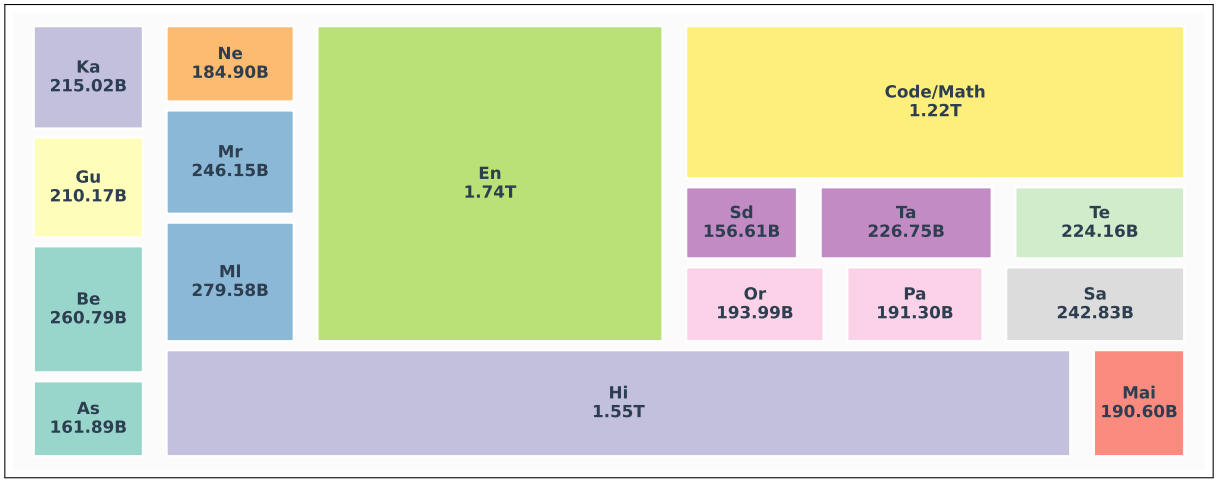


Figure 1 Token Count Distribution

Contents

1 Introduction

Early monolingual models such as BERT [12] and GPT [53, 54, 4] achieved remarkable performance across various NLP tasks, including text classification, question answering, and language modeling. However, these models were primarily trained on English corpora, leaving the vast majority of the world's languages, including many Indic languages, severely underrepresented. Non-English languages had tiny or negligible corpora available for training, which limited the applicability of these models across diverse linguistic communities. This created practical challenges for cross-lingual transfer, zero-shot learning, and domain adaptation, as models could not generalize effectively to languages they had sparsely or never seen. In many real-world applications, this resulted in lower performance for non-English text, and in some cases, models were effectively unusable for low-resource languages. A simple illustration of this limitation is the comparison of language coverage across popular models: BERT and GPT cover very few languages beyond English, whereas models such as mBERT [12] and XLM-R [8] include 100+ languages, highlighting the need for multilingual representation.

The rise of multilingual models and datasets was driven by the recognition that a large portion of the global population was excluded from the benefits of AI due to the heavy reliance on English. Approximately 1.5 billion people speak English worldwide[1], but there are over 7,000 languages globally, with billions of speakers in languages such as Hindi, Bengali, Tamil, and Mandarin. As illustrated in Figure 2, Indic languages collectively represent a substantial portion of the world's linguistic diversity, yet they remain heavily underrepresented in digital datasets. Despite the large number of native speakers, the amount of available digital text for these languages is minuscule compared to high-resource languages. This imbalance highlights a critical global inequity: while Indic languages account for a significant share of the world's spoken languages, their limited digital footprint restricts the accessibility and performance of AI systems for billions of users.

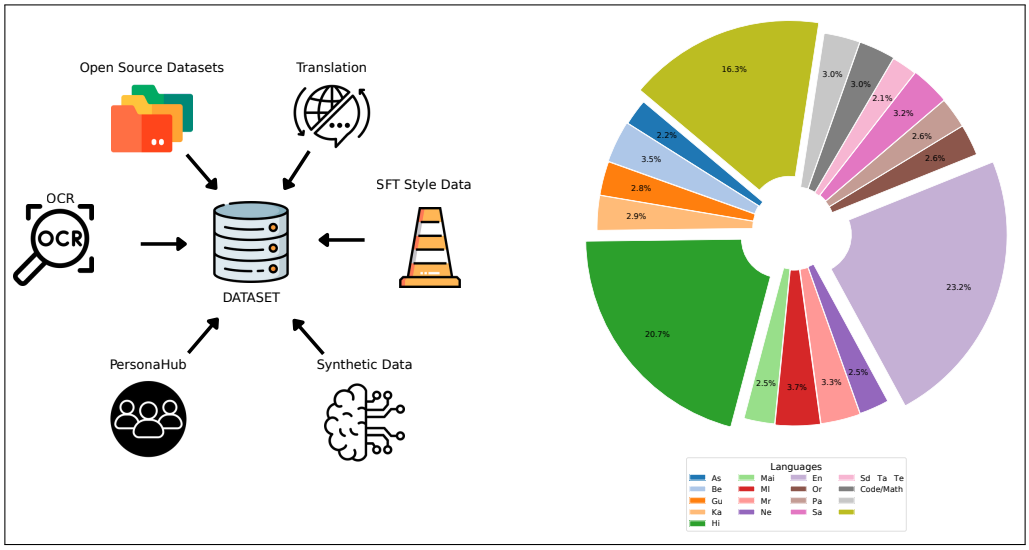


Figure 2 Dataset Sources and Distribution

Within India, the situation is even more pronounced. As depicted in Figure 2, India alone is home to hundreds of languages and dialects, with 16 major languages spoken by tens of millions of people each. Despite this linguistic richness, curated, high-quality digital text for these languages is extremely limited. The scarcity of digital resources affects both model training and evaluation, leading to poorer OCR accuracy, translation errors, and inconsistent NLP performance for Indian languages. These figures underscore the urgent need for large-scale multilingual datasets specifically targeting Indic languages to bridge the gap between speaker population and digital representation.

Multilingual models such as mT5 [76], XLM-R [8], BLOOM [39], LLaMA [70, 71, 21], Gemma [64, 65, 66], Mistral [24], Qwen [2, 77, 52, 15], and Nemotron [42] have expanded NLP capabilities across dozens of languages, relying on large-scale multilingual corpora including Common Crawl [7], Wikipedia dumps [75], CCMatrix [58], mC4 [76], OSCAR [44], and Dolma [62]. However, these models remain biased toward high-resource languages with abundant

[1] https://www.worldometers.info/

training data, while low-resource languages, including most Indic languages, remain underrepresented. Tokenization schemes such as byte-pair encoding or SentencePiece, though effective in high-resource languages, can fragment Indic words into unnatural subwords, further degrading performance. Consequently, despite the existence of multilingual datasets and models, billions of non-English speakers are underserved by AI systems.

Existing data acquisition and preprocessing pipelines, including OCR and translation frameworks, have addressed these challenges only partially. OCR tools struggle with complex scripts such as Devanagari, Tamil, and Malayalam, leading to high error rates in digitized text. Translation pipelines, both machine and human-assisted, often fail to capture the nuances of Indic languages, producing synthetic corpora that are inconsistent or low-quality. Prior attempts to create Indic datasets [27, 57, 13] have either been small in scale or restricted to a limited number of languages, leaving a need for more comprehensive solutions.

Beyond data creation, a critical aspect of this work is the rigorous evaluation of the resulting models. We utilize benchmarks such as Indic MMLU [22] and introduce a parity-based metric to quantify performance disparities across languages. This metric provides an equitable and interpretable measure of model fairness, highlighting both strengths and weaknesses in serving low-resource Indic languages. By evaluating models in this manner, we ensure that the dataset and resulting language models are not only large-scale but also practically useful across diverse languages and domains.

In addition to rigorous evaluation, this work makes several significant contributions that advance the state of Indic NLP. We provide an open-source release of *2 trillion tokens of high-quality pretraining data* spanning 22 Scheduled Indian Languages, accompanied by *300 million image text pairs* to support the development of Indic OCR systems and vision-language models. To facilitate generative modeling and personalization, we release the *Indic PersonaHub*, containing *200 million virtual Indian personas*. We also provide India-centric parallel translated corpora across 22 Scheduled Indian Languages, together with the *Indic MMLU* benchmark covering the same set of languages to enable robust evaluation of multilingual models. Complementing these datasets are domain-specific Indian taxonomies and high-quality web-crawled English corpora, which support cross-lingual learning and transfer.

To support data collection and preparation, we have developed customized crawling and scraping pipelines for Indian web resources, along with the first-of-its-kind synthetic *OCR benchmark, ISOB*, covering 22 Scheduled Indian Languages. This benchmark allows systematic evaluation of OCR performance across diverse scripts and text types. Collectively, these contributions provide a comprehensive ecosystem for Indic NLP, addressing both the scarcity of high-quality data and the lack of robust evaluation frameworks. By combining large-scale corpora, multilingual benchmarks, synthetic datasets, and structured pipelines, this work establishes a foundation for the next generation of robust, equitable, and high-performing Indic language models.

2 Related Work

2.1 English heavy and multilingual corpora

Large scale corpora such as RedPajama ([74]), SlimPajama ([60]), DCLM ([30]), The Pile ([17]), Zyda ([69, 68]), and TxT360 ([63]) have driven major advances in NLP but remain heavily concentrated in English, leaving the majority of the world's languages underrepresented. Multilingual collections (mC4 ([76]), OSCAR ([44]), CC100, ROOTS ([28]), ParaCrawl ([3]), FineWeb2 ([49]), CulturaX ([41]), MultiUN ([14]), Dolma ([62])) 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.

2.2 Indic focused datasets and their limitations

Several recent efforts target Indian languages, but important limitations persist. Parallel collections such as Samanantar [57] and synthetic corpora like Sangraha Synthetic [27] rely on translations or non-native sources, which can produce text lacking cultural authenticity. Monolingual efforts such as IndicCorp [13] 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.

2.3 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 ([29, 26, 82, 59]) have been emphasized in large scale projects ([74, 60, 17]), yet these methods are typically optimized for high resource languages and degrade significantly under the noisy, code mixed conditions common in Indic data ([45]).

OCR remains a significant bottleneck for Indic languages such as Devanagari, Bengali, and Tamil. As highlighted by [35], 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 ([45, 79]). 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.

2.4 Evaluation and parity

Benchmarks such as FLORES ([20]), IndicGenBench ([61]) and MILU ([72]) 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 Indic MMLU

3.1 Motivation

The evaluation of state-of-the-art large language models on Indian languages has long suffered from the absence of a standardized, high-quality benchmark. Existing resources overwhelmingly focus on English, which obscures the question of whether models genuinely understand Indian languages or merely approximate them through translation. To address this gap, we developed the Indic-MMLU benchmark, a multilingual adaptation of the widely used CAIS/MMLU [22] dataset, designed specifically for rigorous evaluation of open-source LLMs across the Indic linguistic landscape. The benchmark encompasses translations into twenty-two Indian languages, of which sixteen have been evaluated in practice alongside English, thereby enabling systematic measurement of knowledge transfer and linguistic generalization across a highly diverse set of languages. The central motivation behind the creation of Indic-MMLU lies in uncovering the degree to which knowledge embedded in English-centric models can be transferred to low-resource Indian languages. If models achieve scores in Indic languages that are comparable to their English counterparts, this provides compelling evidence of cross-lingual generalization and deeper semantic understanding rather than superficial reliance on translation. Conversely, significant disparities highlight the inequities inherent in current training regimes, where high-resource languages dominate representation and low-resource languages remain marginalized. By making this gap visible through a rigorously constructed benchmark, Indic-MMLU provides both a tool for evaluation and a signal for future data and model development.

3.2 Benchmark Construction

The construction of the benchmark followed a carefully designed pipeline that integrated automated methods, large-scale language models, and human expertise. Starting with the original English MMLU dataset, translations were generated into each target Indic language using strong machine translation systems such as IndicTrans2 [16]. However, rather than accepting raw machine outputs, we enhanced these translations using large-scale LLMs guided by task-specific instructions, ensuring that phrasing, mathematical expressions, and answer choice alignments were preserved with fidelity. To verify semantic consistency with the English source, we embedded both original and translated instances into a shared embedding space and measured cosine similarity, filtering or revising translations that exhibited low alignment. This automated process was complemented by human-in-the-loop evaluations, in which native

speakers, linguists, and subject matter experts rated translations for fluency, correctness, and coherence. We further introduced the use of LLM-as-judge personas, where specialized evaluation prompts simulated linguistic experts, mathematical experts, and coherence judges to provide scalable and reproducible assessments. Back-translation, embedding similarity, and traditional n-gram metrics such as BLEU, ROUGE, and ChrF++ [46, 31, 50] were employed to provide quantitative measures of translation fidelity. This multi-stage validation process ensured that the resulting benchmark was both linguistically natural and semantically faithful to the original.

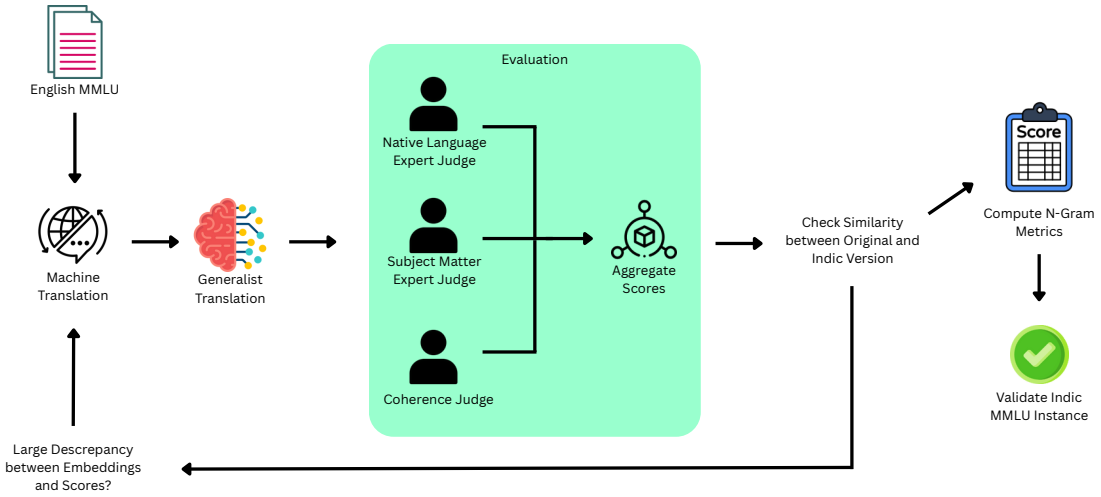


Figure 3 Indic MMLU Creation Workflow

To scale evaluation of translation quality, we employed LLMs as judges under three distinct expert personas: *Math Expert*, *Linguistic Expert*, and *Coherence Expert*. Each persona rated translations on a 1–10 scale, measuring mathematical fidelity, fluency and idiomaticity, and semantic coherence, respectively. Table 2 summarizes the average ratings across Indic languages.

Table 1 Cosine similarity scores between English source and Indic translations.

Lang	Mean	Std	Min	Max
Assamese	0.8045	0.0597	0.4644	1.0000
Bengali	0.8133	0.0504	0.5276	1.0000
Gujarati	0.8211	0.0535	0.4016	1.0000
Hindi	0.8472	0.0489	0.3688	1.0000
Kannada	0.8106	0.0565	0.4660	1.0000
Maithili	0.8226	0.0504	0.4936	1.0000
Malayalam	0.8158	0.0531	0.5109	1.0000
Marathi	0.8129	0.0513	0.5281	1.0000
Nepali	0.8242	0.0502	0.4802	1.0000
Oriya	0.8159	0.0555	0.4684	1.0000
Punjabi	0.8246	0.0516	0.5450	1.0000
Sanskrit	0.7912	0.0574	0.4981	1.0000
Sindhi	0.7646	0.0735	0.3633	0.9676
Tamil	0.7964	0.0559	0.5242	1.0000
Telugu	0.8006	0.0524	0.5379	1.0000

Table 2 Summary of LLM-as-Judge Ratings across Indic languages. Ratings are on a 1–10 scale.

Lang	Maths	Coh	Linguist	Records
Assamese	9.15	9.04	8.13	14042
Bengali	9.18	9.13	8.20	14042
Gujarati	9.26	9.21	8.37	14042
Hindi	9.65	9.72	8.66	14042
Kannada	9.32	9.11	8.13	14042
Maithili	9.26	9.12	8.22	14042
Malayalam	9.13	8.99	7.96	14042
Marathi	9.30	9.34	8.38	14042
Nepali	9.47	9.33	8.40	14042
Oriya	9.26	8.94	8.22	14042
Punjabi	9.35	9.05	8.25	14042
Sanskrit	8.34	8.11	6.39	14042
Sindhi	2.07	3.11	2.48	14042
Tamil	9.05	9.04	8.07	14042
Telugu	9.22	8.98	7.99	14042

In addition to qualitative ratings, cosine similarity between embeddings of English source questions and their Indic translations provided a scalable quantitative check of semantic faithfulness. Table 1 reports mean, standard deviation, and range of similarity values across languages, with most scores clustering around 0.80–0.85, indicating strong semantic alignment.

3.3 Experiments and Evaluation

With the benchmark constructed, we conducted one of the most extensive evaluations of open-source language models on Indic languages to date. Using standardized frameworks such as lm-eval, approximately two dozen models including mid-sized multilingual transformers and cutting-edge open-source LLMs were evaluated on Indic-MMLU across 16 Indic languages and English. Performance was measured using exact-match accuracy, and results were aggregated at both per-language and cross-language levels.

Table 3 Absolute Indic MMLU performance

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**

Evaluation Prompt: Coherence

Context

You are a Coherence Expert with academic grounding in comparative linguistics, logic, and translation studies, specializing in assessing semantic alignment across languages. They evaluate whether the translated text maintains the intended meaning, logical flow, and clarity of the original English. They look for shifts in nuance, omissions, or distortions that could affect comprehension. Their approach is holistic: ensuring that the reader of the translated version understands the same idea as the reader of the English original. They provide a rating from 1 to 10 for overall coherence between the two texts.

Objective

Evaluate the Translated Question for linguistic quality only:
Focus only on semantic and logical coherence: - Does the translated text preserve the same meaning and intent? - Is the flow and structure aligned with the original? - Does the translation stay faithful without distortion or loss?

Inputs:

Original Question (English):
`"{0}"`
`"{1}"`

Translated Question:
Language of Translation: `"{3}"`
`"{2}"`

Output Format (strict JSON):

Give a single integer rating from 1 to 10 (1 = worst, 10 = excellent).

```
{
  "Coherence Expert": <rating>
}
```

Our evaluations revealed a consistent performance gap between English and Indic languages. While strong models such as DeepSeek and Gemma achieved relatively robust results across multiple Indic languages, their performance remained substantially below their English benchmarks. Languages with greater digital presence, such as Hindi and Bengali, benefited from higher accuracy, while those with scarce representation in digital corpora, such as Sindhi and Sanskrit, suffered severe performance drops. Languages with greater digital presence, such as Hindi and Bengali, benefited from higher accuracy, while those with scarce representation in digital corpora, such as Sindhi and Sanskrit, suffered severe performance drops.

Evaluation Prompt: Linguist

Context

You are a Linguistic Expert highly trained in "3" grammar, vocabulary, stylistics, and cultural context, with professional experience in translation and language pedagogy. They evaluate whether the translation reads naturally, adheres to formal linguistic rules, and avoids machine-like literalness or awkward phrasing. They pay attention to register, idiomatic usage, and fluency, ensuring the translation would feel authentic to a native speaker. Their judgments balance linguistic purity and readability, assigning a rating from 1 to 10 for linguistic quality and naturalness.

Objective

Evaluate the Translated Question for linguistic quality only:
Is grammar correct and natural? Is the vocabulary appropriate and idiomatic for the target language? Does the sentence sound fluent and well-structured?

Inputs:

Original Question (English):
`"{0}"`
`"{1}"`

Translated Question:
Language of Translation: `"{3}"`
`"{2}"`

Output Format (strict JSON):

Give a single integer rating from 1 to 10 (1 = worst, 10 = excellent).

```
{
  "Linguist Expert": <rating>
}
```

Languages with greater digital presence, such as Hindi and Bengali, benefited from higher accuracy, while those with scarce representation in digital corpora, such as Sindhi and Sanskrit, suffered severe performance drops. Radar plots and heatmaps of performance further emphasized this disparity, underscoring both the progress achieved in high-resource Indic languages and the persistent challenges in low-resource contexts. These findings illustrate the double-edged nature of multilingual LLMs: while they demonstrate encouraging signs of cross-lingual transfer, their benefits remain unevenly distributed across Indic languages. Indic-MMLU thus not only quantifies current performance levels but also provides a roadmap for targeted interventions, improving data quality, expanding coverage for underrepresented languages, and advancing tokenization strategies better suited for Indic scripts. By openly releasing both the dataset and evaluation protocols, Indic-MMLU establishes a foundation for reproducible research and equitable benchmarking in Indic NLP.

4 Data Acquisition

Building equitable representation for India's 22 official languages demands more than incremental improvements to existing pipelines. It requires rethinking data acquisition from the ground up. The challenge is not simply one of scale,

but of diversity, authenticity, and cultural grounding. Most large language models today are trained on web-scraped corpora that privilege high-resource languages, leaving Indic languages with a fraction of the data volume and virtually none of the domain-specific depth needed for real-world applications.

The foundation of MILA draws from established multilingual corpora similar to Pile [17], RedPajama [74], and C4 [56], incorporating approximately 5 billion words gathered through multi-source web crawling of multilingual websites, forums, and academic repositories, apart from over 1700 open datasets from HuggingFace[2]. However, web-scraped data alone cannot address the cultural and linguistic gaps inherent in translated or synthetic corpora. To mitigate these limitations, we prioritized curated book collections, amounting to approximately 32 billion words across 16 languages, sourced primarily from Archive.org[3] and the National Digital Library of India[4]. These collections offer authentic content and culturally grounded materials that complement the breadth of crawled and open-source data, ensuring that MILA captures not just linguistic patterns but also the educational and cultural contexts that define how these languages are actually used.

Collecting these books and academic papers posed significant systems challenges that extended far beyond simple download automation. Standard pipelines frequently stalled, exceeded resource limits, or required weeks of wall-clock time, making large-scale acquisition impractical without fundamental architectural innovations. The heterogeneity of sources demanded source specific solutions that could operate at unprecedented concurrency while maintaining provenance tracking, license compliance, and fault tolerance. We therefore designed custom infrastructure and optimization layers that reduced runtime and compute usage by 40–70%, enabling sustained acquisition at scales that standard tools could not achieve. This section details our approach across four major acquisition domains: Archive.org, NDLI, Wikimedia, and the underlying infrastructure that made large-scale harvesting feasible.

4.1 Archive.org

Archive.org represents one of the most valuable yet underutilized resources for Indic language modeling. Unlike web-scraped text, which often suffers from quality inconsistencies, duplication, and cultural decontextualization, Archive.org provides digitized books that have undergone editorial processes, domain specialization, and cultural embedding. For Indic languages in particular, Archive.org hosts extensive collections of historical texts, classical literature, government documents, and subject-specific materials that are virtually absent from standard web corpora. However, accessing this wealth of content at scale required overcoming substantial technical and organizational challenges related to metadata discovery, download orchestration, and quality assurance.

Table 4 Archive.org 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 (Maths, Ayurveda)	45.10 K	2.57 M	0.89 B
Total	**1 M**	**84.00 M**	**27.15 B**

As seen in Table 4, Archive.org, we extracted over 1,012,198 digitized PDFs spanning multiple Indic languages and government documents, totaling 82,104,639 pages. This collection forms one of the richest curated corpora for Indic language model training. Bengali and Sanskrit provide substantial depth, contributing 10.95 million and 10.09 million

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

pages, respectively, encompassing 3.10 billion and 2.68 billion words. Bengali materials include both modern literature and historical documents from the Bengal Renaissance, while Sanskrit holdings feature philosophical treatises, grammatical texts, and mathematical manuscripts spanning millennia. Hindi offers the highest number of individual PDFs at 396,120, with 7.53 million pages and 4.15 billion words, reflecting contemporary textbooks, popular literature, and government publications. Malayalam and Marathi further enhance the corpus, with Malayalam's 65,030 PDFs spanning 2.18 million pages (1.06 billion words) and Marathi's 124,220 PDFs covering 3.02 million pages (1.26 billion words), providing a mix of traditional, historical, and contemporary literary forms. Government documents across multiple languages supplement these sources, offering structured policy-oriented content, including legal codes, census reports, and administrative guidelines, grounding language models in governance-specific discourse.

To enhance subject coverage beyond general literature and government documents, we further targeted subject collections such as Mathematics, Ayurveda, and Agriculture. These domain-focused subsets represent critical areas where Indic-language expertise exists but is poorly represented in standard training corpora. Mathematical texts include both classical Indian mathematics such as works on algebra, geometry, and astronomy from historical mathematicians, and modern textbooks covering calculus, statistics, and applied mathematics. Ayurveda collections encompass classical texts, commentaries, materia medica, and clinical guidelines, providing comprehensive coverage of traditional medical knowledge. Agriculture materials include crop management guides, soil science texts, veterinary manuals, and rural development documents that reflect India's agricultural diversity and traditional farming knowledge. These domain-focused subsets were subjected to optical character recognition post-processing and correction via large language models, improving text quality and making the data more usable for supervised fine-tuning.

The acquisition process for Archive.org materials required navigating several technical challenges. Archive.org exposes its metadata through paginated search endpoints, each capped at approximately 10,000 results. For large collections spanning hundreds of thousands of items, this pagination limit necessitates sophisticated query strategies that partition the search space into manageable chunks. Metadata quality is heterogeneous across collections, with inconsistencies in language tagging, missing bibliographic information, and ambiguous licensing statements. Long-running download jobs are fragile, often stalling due to network issues, rate limiting, or server-side errors. Naïve sequential ingestion approaches required several days even on high-bandwidth machines, making comprehensive collection infeasible within reasonable time frames. The details on how these challenges were tackled are given in 4.4.

4.2 NDLI

Table 5 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 6 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

The National Digital Library of India represents a fundamentally different acquisition paradigm compared to Archive.org.

Where Archive.org provides broad historical and literary coverage, NDLI offers curriculum-aligned, licensed, and provider-attributed materials explicitly designed for educational use. This structural difference makes NDLI uniquely valuable for training language models that must operate in pedagogical contexts, answer curriculum-based questions, or generate educational content aligned with Indian educational standards. NDLI materials span two disjoint strata: school and state-boards covering K-12 education, and higher education encompassing undergraduate, postgraduate, and diploma programs. We report each stratum on its own axes to avoid conflating counts across heterogeneous groupings, allowing us to preserve the pedagogical and institutional context of each item while facilitating targeted OCR and post-correction workflows adapted to Indic scripts and educational content.

State-board and NCERT materials provide grade-sequenced, syllabus-aligned content in Indic languages and English, making this stratum essential for curriculum-grounded pretraining and supervised fine-tuning on pedagogy-aligned tasks such as worked solutions, syllabus-based question answering, and instructional content generation. The curriculum alignment is not merely topical but structural materials follow prescribed syllabi, use standardized terminology, and progress through concepts in pedagogically validated sequences. This makes NDLI school content particularly valuable for applications like automated tutoring systems, homework assistance, and educational content generation that must respect both subject matter and grade-appropriate presentation.

Table 7 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

The distribution of NDLI school and state-board content reveals the linguistic and institutional landscape of Indian education, as detailed in Tables 5, 6, and 7. As shown in Table 5, Hindi dominates with 4,114 items, reflecting both its status as a widely taught language and the extensive digitization efforts by Hindi-medium state boards. English follows closely with 3,785 items, representing both English-medium schools and English as a subject across state boards.On te other hand, several smaller languages including Garo with 10 items, Bodo with 4 items, Nepali with 1 item, and Manipuri with 1 item highlight the uneven digitization across states, with implications for equitable language model development.

The distribution by class level, presented in Table 6, shows relatively balanced coverage across grades with some concentration in secondary and higher secondary levels. Class X leads with 1,926 items, reflecting the significance of board examinations at this level and corresponding digitization priority. Class XI follows with 1,611 items, Class IX with 1,533 items, and Class XII with 1,483 items. These four grades collectively account for the bulk of materials, corresponding to the secondary education phase where standardized curricula are most rigorously defined and assessment is most formal. Middle and primary levels are also represented, though less prominently, ensuring coverage across the full curricular progression. This stratification supports the design of training pipelines that respect pedagogical sequencing and enables creation of evaluation sets that test a model's ability to generate grade-appropriate explanations

without oversimplification or unnecessary complexity.

The provider distribution, presented in Table 7, highlights both the institutional diversity of the collection and the uneven levels of digitization commitment across Indian states. SCERT Telangana leads with 4,247 items, reflecting the state's strong investment in digital curriculum resources as part of recent education initiatives. Raj-eGyan (Rajasthan) and Punjab School Education Board follow with 2,606 and 2,465 items respectively, underscoring the momentum of state-led digitization efforts in northern India. Other contributors such as Gujarat, Jammu and Kashmir, and NCERT provide substantial but comparatively smaller shares, while several states and institutions add more modest numbers. Collectively, the distribution illustrates that while some states have achieved large-scale digitization, others remain underrepresented, pointing to regional disparities in access to digital educational content.

These tables provided complementary perspectives on the same corpus: languages highlight linguistic diversity and the multilingual nature of Indian education, grade levels ground pedagogy and enable curriculum-sequenced model training, and providers reveal provenance and institutional commitment to open education. 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 ensures that a single mathematics textbook available in both Hindi and English, or a multi-grade resource spanning Classes IX and X, is counted once but tagged with all applicable metadata. This approach enables flexible querying and sampling strategies during training while preventing artificial inflation of dataset statistics.

Table 8 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 9 NDLI Higher Education: counts by education level (items).

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

Table 10 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

Higher education holdings from NDLI provide domain depth in Mathematics, Botany, Chemistry, Medicine, and Engineering, along with provider-curated texts amenable to precision supervised fine-tuning on derivations, definitions, proofs, and technical procedures. These materials differ fundamentally from school content in their assumed prior knowledge, technical depth, and specialized vocabulary. Higher education content is valuable not just for training models to understand advanced topics but also for enabling retrieval over structured knowledge, where precise definitions, formal proofs, and established methodologies must be accurately represented and retrievable.

The provider distribution for higher education, shown in Table 8, is dominated by LibreTexts with 7,591 items, reflecting its multi-institutional effort to curate open educational resources across STEM disciplines. e-Adhyayan follows closely with 6,902 items, highlighting a major Indian initiative aligned with national curricula. Other contributors include the Botanical Survey of India with 976 items, offering authoritative botanical references, and e-KUMBH with

478 items, which expands access to higher education materials in Indian languages. The distribution by education level, presented in Table 9, shows a clear emphasis on postgraduate content with 6,901 items, followed by 1,259 undergraduate and 146 diploma-level resources. This concentration at the postgraduate level underscores the advanced and specialized nature of the collection, making it particularly valuable for training models on technical reasoning, research-oriented writing, and domain-specific knowledge.

The disciplinary specialization, shown in Table 10, is led by Mathematics with 2,884 items, offering extensive coverage across core and advanced topics that are especially valuable for developing models with strong reasoning capabilities. Botany follows with 900 items, reflecting India's rich biodiversity and the strong institutional contributions in this field. Commerce and related areas contribute 875 items, underscoring the practical relevance of economic and infrastructural studies. Other disciplines such as chemistry, medicine, and engineering add further breadth, ensuring the corpus is not only quantitatively rich but also balanced across technical, scientific, and applied domains. This subject diversity, when considered alongside provider and level distributions, enables the construction of evaluation-ready subsets tailored to curriculum progression and disciplinary expertise.

The NDLI pipeline illustrates a methodology that extends beyond mere content acquisition to systematic organization that preserves pedagogical and institutional context. Unlike Archive.org's historical focus or Wikimedia's encyclopedic coverage, NDLI provides materials explicitly designed for learning, with clear curricular alignment, grade-level appropriateness, and institutional provenance. This makes NDLI content particularly valuable for educational applications of language models, where generating pedagogically sound content, respecting curricular sequences, and providing grade-appropriate explanations are critical requirements that web-scraped data cannot reliably support.

4.3 Wikimedia

Wikimedia[5] projects form one of the largest open-access multilingual resources for Indian languages, capturing encyclopedic, cultural, educational, and archival text that complements the historical depth of Archive.org and the curricular structure of NDLI. Where Archive.org provides edited books and NDLI offers curriculum-aligned materials, Wikimedia contributes community-maintained, collaboratively edited content that reflects contemporary knowledge, cultural perspectives, and living linguistic practices. The distributed nature of Wikimedia projects—spanning Wikipedia, Wikisource, Wikibooks, and numerous specialized initiatives—provides diverse textual genres and knowledge domains, making it an essential component of comprehensive multilingual training corpora.

Table 11 Wikimedia Corpora: Language-Wise Summary

Language	Words (in Millions)	Files
English	2.80B	3.65M
Bengali	383.53M	996K
Hindi	159.89M	355K
Tamil	152.45M	681K
Telugu	108.38M	261K
Urdu	103.67M	168K
Sanskrit	89.76M	187K
Malayalam	99.08M	326K
Gujarati	25.88M	70K
Marathi	54.13M	149K
Kannada	52.87M	121K
Oriya/Odia	18.52M	61K
Punjabi	40.48M	112K
Assamese	24.15M	74K
Kashmiri	1.53M	3.9K
Nepali	21.20M	43K
Others (Manipuri, Garo, Bodo, etc.)	< 1M	< 2K

The language-wise distribution of Wikimedia content, presented in Table 11, reveals both the platform's multilingual

[5] https://commons.wikimedia.org/

breadth and the persistent resource imbalances across languages. English dominates overwhelmingly with 2.80 billion words across 3.65 million files, reflecting Wikipedia's origins as an English-language project and the continued predominance of English in online knowledge production. This massive English presence, while valuable for multilingual models that must handle code-switching and cross-lingual tasks, also highlights the scale of resource disparity that MILA aims to address.

Bengali emerges as the strongest Indic language with nearly 384 million words, supported by active editor communities and institutional digitization initiatives that have enabled systematic content creation. Hindi and Tamil follow with sizable volumes, though Hindi's output remains modest relative to its vast speaker base, illustrating the persistent digital divide even among widely spoken languages. Mid-resource languages such as Telugu, Malayalam, Marathi, Kannada, and Punjabi contribute substantial content, yet still represent only a fraction of the available English corpus. At the other end of the spectrum, languages like Gujarati, Assamese, Nepali, and Oriya remain underrepresented, while Kashmiri and several smaller languages, including Manipuri, Garo, and Bodo, contribute only marginal volumes. This stark disparity between high- and low-resource Indic languages underscores the uneven digital landscape and highlights the urgent need for targeted curation efforts to bridge linguistic inequities.

This language-wise view emphasizes both the relative strengths of English and high-resource Indic languages and the severe under-representation of low-resource languages. The long-tail distribution has profound implications for language model training: while English and Bengali content can support robust monolingual models, languages like Kashmiri, Manipuri, and Bodo require cross-lingual transfer, synthetic augmentation, and careful low-resource techniques to achieve even basic competence. The stark disparities also underscore why curated collections from Archive.org and NDLI are essential, web-crawled and community-maintained sources alone cannot provide the volume and quality needed for equitable language modeling across all Indic languages.

Table 12 Wikimedia Corpora: Project-Wise Summary

Project	Words (in Billions)	Files (in Millions)
Wikisource	2.09B	4.95M
Wikipedia	0.15B	0.44M
Wikibooks	0.10B	0.09M
Wikiquote	0.099B	0.07M
Wikinews	0.015B	0.03M
Wikiversity	0.051B	0.04M
Wikivoyage	0.044B	0.03M
Others (Kidatawiki, Iwiki, Ecieswiki)	0.17B	1.1M
Grand Total	**4.42B**	**8.93M**

The project-wise distribution, summarized in Table 12, reveals how Wikimedia content is distributed across different knowledge domains and textual genres. Wikisource dominates with 2.09 billion words across 4.95 million files, providing archival and historical texts. Wikisource's mission is to collect and transcribe source documents—original texts, historical documents, literary works, and primary sources, that are in the public domain or permissively licensed. For Indic languages, Wikisource is particularly valuable because it hosts digitized versions of classical literature, historical chronicles, religious texts, and early modern works that are often unavailable in other digital formats. The dominance of Wikisource in total word count reflects both the length of these source documents and the systematic digitization efforts by language communities.

The Wikimedia ecosystem contributes a total of 4.42 billion words across 8.93 million files, forming one of the most diverse open repositories for multilingual content. Wikisource dominates in scale due to its digitized literary and historical texts, while Wikipedia, despite contributing fewer words, provides unparalleled topical breadth, structured knowledge, and contemporary relevance. Educational projects such as Wikibooks and Wikiversity add structured pedagogical materials, offering valuable resources for instructional and fine-tuning tasks. Smaller projects like Wikiquote, Wikinews, and Wikivoyage, though modest in size, enrich the corpus with idiomatic expressions, journalistic writing, cultural context, and descriptive language. Collectively, these repositories complement one another: literary depth from Wikisource, encyclopedic coverage from Wikipedia, didactic clarity from Wikibooks and Wikiversity, and domain-specific perspectives from smaller projects; ensuring broad linguistic and thematic diversity for model training.

The distribution highlights both the strengths and limitations of Wikimedia's community-maintained content. Wikisource contributes archival text with temporal depth and literary richness, while Wikipedia offers encyclopedic coverage and structured factual grounding. Together they support historical and contemporary linguistic research as well as information-seeking and question-answering tasks. Smaller projects such as Wikibooks, Wikiquote, Wikinews, Wikiversity, and Wikivoyage add genre diversity, exposing models to instructional writing, quotations, journalism, and travel description. At the same time, coverage is highly uneven: English and a handful of Indic languages dominate with billions of words, while many others remain under-represented. This imbalance necessitates complementary resources—curated books from Archive.org, curriculum materials from NDLI, and synthetic augmentation through the Indic-Persona Hub (Section 8). Wikimedia alone, though valuable, cannot provide sufficient representation for low-resource languages or the domain depth needed for specialized areas like agriculture, Ayurveda, or law.

Quality also varies substantially across languages. High-resource languages benefit from active editor communities, established guidelines, and systematic patrolling, while low-resource languages often face smaller communities, inconsistent editing, and greater vulnerability to low-quality contributions. To address this, we apply language-specific quality filtering based on article length, structural completeness, reference density, and community quality markers such as featured or good article status. In combination, Wikimedia's contemporary breadth and community perspectives, Archive.org's historical depth, and NDLI's institutional framing provide MILA with the linguistic diversity, domain coverage, temporal range, and cultural authenticity needed for equitable multilingual language modeling across India's diverse linguistic landscape.

4.4 Infrastructure and Optimisation

Beyond the content and organizational strategies detailed in the previous sections, the acquisition of MILA required fundamental innovations in systems infrastructure and optimization. Collecting large-scale academic corpora from heterogeneous sources such as Archive.org and NDLI was not merely a data challenge but also a systems challenge demanding custom-built solutions that could operate at unprecedented scale while maintaining reliability, provenance, and efficiency. Standard pipelines frequently stalled, exceeded resource limits, or required weeks of wall-clock time, making comprehensive acquisition impractical. We therefore designed custom infrastructure and optimization layers that reduced runtime and compute usage by 40–70%, enabling sustained acquisition at scales that generic tools could not achieve. This section details the technical architecture, optimization strategies, and governance frameworks that made large-scale multilingual corpus construction feasible.

Table 13 BitTorrent performance comparison: standard clients vs. our optimized engine.

Metric	Standard BT	Our Engine	Improvement
Zero-leech speed	~50 KB/s	~200 KB/s	3×
Max connections	~200	30,000	150×
Cache usage	~5% RAM	60% RAM	12×
Concurrent downloads	3–5	30	6×
Stall recovery	Manual	Auto (5–10s)	—

Large-Scale Corpus Acquisition: A Metadata-First Architecture. Large-scale corpus acquisition presents a fundamental trade-off: high throughput demands aggressive parallelism but increases failures and complicates provenance tracking, while reliability requires conservative allocation and checkpointing that reduces throughput. Governance adds further overhead through metadata tracking, license verification, and audit trails. We resolve this through a metadata-first architecture that treats metadata as the primary object for discovery, governance, and deduplication before text processing. This enables license-aware filtering before download, record-level deduplication via stable identifiers, and targeted discovery by subject, language, and education level, while reducing costs by eliminating redundant downloads.

Unified Metadata Schema and Hierarchical Deduplication. At the scale of tens of millions of documents, direct text-based deduplication or OCR is infeasible as a first pass. We therefore construct a unified metadata schema that normalizes identifiers (ISBN, DOI, handle, archive ID), bibliographic data (title, author, publisher, year, edition), technical attributes (filesize, extension, pages), governance information (license, rights, timestamps), and integrity checks (MD5, URL, cover image). This consistent representation enables license-aware filtering and prepares the

ground for robust deduplication. Deduplication proceeds hierarchically by combining multiple signals to maximize accuracy while minimizing false positives. Records are marked as duplicates if any hard key—canonical URL, DOI, ISBN, or MD5—matches exactly. Otherwise, soft matching bundles normalized title, author, and publication year, with corroborating attributes like filesize and page count, to identify duplicates within tolerance thresholds. All decisions are logged with record IDs and triggering signals for auditability. This pipeline effectively eliminates redundancies, particularly in Archive.org where popular books often reappear across collections, mirrors, and formats (PDF, EPUB, DJVU).

Governance-First Licensing. Since many collections contain restrictive or ambiguous licenses, we enforce governance as a primary constraint. Each item is tagged with licensing metadata and provenance snapshots. Non-permissive or uncertain items are quarantined for research-only use, contributing to coverage statistics but excluded from training and redistribution. Only permissively licensed content from Archive.org public domain collections, NDLI open textbooks, and Wikimedia projects enters the training pipeline.

Source-Specific Challenges. Archive.org provides rich bibliographic metadata including stable identifiers, collection memberships, and language tags, but exhibits occasional language field errors with Hindi mislabeled or Indic content lacking tags. Technical challenges include paginated search endpoints capped at 10,000 results per query and fragile long-running jobs. We addressed this via adaptive query planning that slices queries by date ranges, subjects, and languages to bypass result windows, feeding asynchronous multi-semaphore crawlers with pause-resume checkpoints. Separate semaphores for metadata fetching, downloading, and post-processing allow each stage to proceed at its natural rate. This reduced ingestion time from 7 days to 24 hours on comparable hardware, saving 40–70% in compute overhead. NDLI supplies structured metadata tied to education levels and subject facets, enabling curriculum-aligned slicing. Challenges include multilingual misclassification, sparse ISBN coverage, and inconsistent subject tagging. We apply normalization layers mapping provider vocabularies to standardized taxonomies and flag ambiguous records for manual review.

Metadata-first processing has documented limitations: pervasive language misclassification (Hindi as English, script-metadata mismatches), licensing gaps requiring collection-based heuristics, and edition ambiguities from multiple ISBNs or minor reprints. Future iterations will integrate OCR-based post-processing: text-based deduplication using MinHash/SimHash to detect edition-level reprints; OCR-based language identification with script-aware classifiers (fastText, CLD3) to correct mislabels; and confidence scoring combining metadata with OCR predictions to quantify uncertainty. Yet, this strategy provides a governed, efficient, and auditable baseline enabling acquisition at scales content-first approaches cannot achieve. While OCR-based deduplication and language identification will further improve quality, this hybrid approach (metadata for scale and governance, content analysis for quality) maximizes the value of large collections for Indic-focused LLM pretraining and domain-specific fine-tuning. The infrastructure represents a foundational contribution extending beyond our immediate needs, providing reusable patterns for equitable language technology development across linguistic diversity.

5 Data Curation

5.1 Pipeline Architecture and Quality Assessment

High-quality training data constitutes the foundation of robust language models. Low-quality or misclassified text degrades performance and introduces undesirable behaviors, a challenge extensively documented in prior work [48, 33, 78, 55]. While existing curation pipelines prove effective for widely studied languages like English, they often fall short for Indian languages, each presenting unique challenges in morphology, script variation, and code-mixing patterns. We address this gap through a multi-stage curation framework that combines NVIDIA NeMo Curator[6] with custom scripts tailored for multilingual Indian corpora, producing a dataset that is simultaneously clean, diverse, safe, and legally compliant.

Our curation pipeline begins with document-level quality assessment using in-house fastText-based classifiers trained for all target languages. This initial classification partitions the corpus into three tiers: high-quality documents that proceed directly to downstream stages, and medium- and low-quality documents routed to a synthetic rewriting pipeline that transforms them into high-quality text while preserving linguistic characteristics, as detailed in Section 8. Following quality stratification, the pipeline applies a series of modifiers and heuristic filters enumerated in Table 14. Mod-

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

Table 14 Comparison of Modifiers and Heuristic Filters.

Modifiers	Heuristic Filters
Boilerplate String Modifier	Word Count Filter
HTML Tag Modifier	Repeating Top NGrams Filter $n = 2$, $n = 3$
Unicode Reformatter	URLs Filter
Quotation Unifier	Symbols to Words Filter
Excess White Space Remover	Numbers Filter

ifiers perform structural normalization: the Boilerplate String Modifier removes templated web artifacts, the HTML Tag Modifier strips markup, the Unicode Reformatter ensures textual consistency, the Quotation Unifier standardizes quote characters, and the Excess White Space Remover eliminates formatting irregularities. Heuristic filters then eliminate degenerate content through Word Count Filters that remove extremely short or long documents, Repeating Top NGrams Filters ($n = 2$, $n = 3$) that detect mechanically generated or spam text, and specialized filters for URLs, excessive symbols, and numeric-heavy content that lack linguistic value.

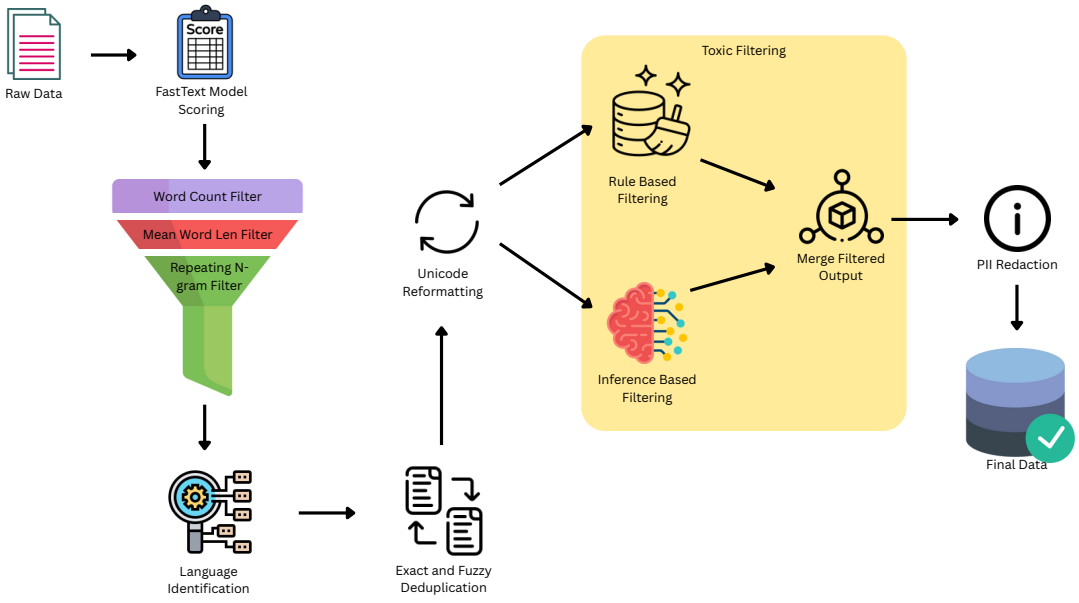


Figure 4 Curation Pipeline

Deduplication operates at two levels of granularity. Exact deduplication removes bitwise identical documents, a straightforward but essential step given the redundancy inherent in web-scale corpora. More critically, GPU-accelerated fuzzy deduplication eliminates near-duplicates arising from template variations, pagination artifacts, and minor textual perturbations common in large-scale web scraping [29, 26]. This aggressive deduplication strategy prevents models from overfitting to repetitive content while substantially reducing training corpus size and associated computational costs. Following Mendu et al. [36], we implement a two-stage toxic filtering process designed to balance safety and recall. Initial rule-based filters flag documents containing explicit slurs, hate speech patterns, and prohibited content categories. Recognizing that rule-based approaches produce false positives—particularly in multilingual contexts where lexical ambiguity and cultural nuance complicate simple pattern matching—we apply multilingual RoBERTa-based [34] inference to reclassify flagged documents, restoring safe content incorrectly captured by heuristics. Finally, the PiiModifier detects and redacts personally identifiable information including names, postal addresses, email addresses, and phone numbers, ensuring privacy compliance across all retained documents.

The complete workflow is visualized in Figure 4, which illustrates the sequential processing stages from raw text ingestion through quality scoring, language identification, filtering, deduplication, toxic content removal, and PII redac-

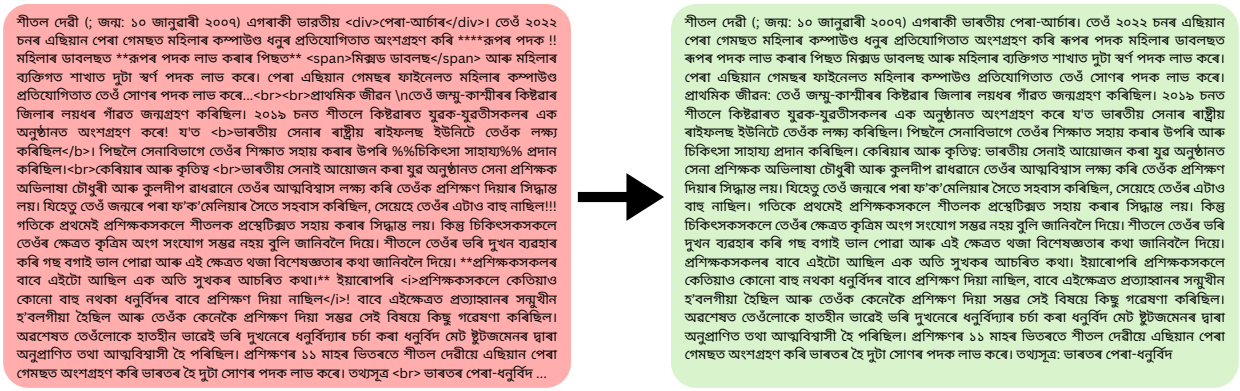


Figure 5 Conventional vs Curated Comparison

tion. To demonstrate the practical impact of these transformations, Figure 5 presents an Assamese text extracted from Common Crawl before and after curation. The raw text exhibits inline HTML tags, formatting artifacts, repeated symbols, and inconsistent spacing characteristic of web-scraped content. Post-curation, these artifacts are systematically removed, yielding clean, readable text that preserves all semantic and contextual information while dramatically improving linguistic fidelity. This example underscores how our approach enhances text quality for low-resource Indic languages where conventional pipelines often fail due to script-specific challenges and limited training data for quality classifiers.

5.2 Ablation Study: Task Performance Improvements

To validate the effectiveness of our curation pipeline, we conduct two complementary ablation experiments that isolate the impact of data quality on model performance and safety. The first experiment directly compares models trained on conventional versus curated data under otherwise identical conditions. We continually pretrain two instances of Param-1 [51], a 2.9 billion parameter causal language model, on 2 trillion tokens of English and Hindi data: one using raw web-scraped text processed only with basic cleaning, and another using the fully curated pipeline described above. The model architecture, detailed in Table 19, employs grouped-query attention with 32 hidden layers, a hidden dimension of 2048, an intermediate dimension of 7168, and fast-swiglu activation functions. All hyperparameters, training duration, batch size, learning rate schedule, and computational infrastructure remain strictly identical between the two experiments, ensuring that any performance differences arise solely from data quality rather than confounding factors.

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

The results, presented in Table 16, demonstrate substantial and consistent improvements across all evaluated benchmarks. On ARC Challenge [38], the curated model achieves 53.6% accuracy compared to 46.5% for the conventional baseline, representing a 7.1 percentage point gain on this challenging reasoning benchmark. ARC Easy [38] shows a

more modest but still meaningful improvement from 73.6% to 74.2%. HellaSwag [80] performance remains stable at approximately 73.5-73.8% for English, but the Hindi variant reveals dramatic gains: the curated model achieves 41.4% accuracy versus only 28.9% for the conventional baseline, a 12.5 percentage point improvement that highlights the particular value of curation for low-resource languages where noisy training data disproportionately degrades performance. MMLU [22] results follow a similar pattern, with the curated model reaching 46.2% on English MMLU compared to 41.3% for the baseline, and 34.6% on Hindi MMLU versus 26.2%, an 8.4 percentage point improvement that again underscores curation's amplified benefits for Indic languages. These results provide compelling evidence that systematic data curation translates directly into stronger downstream task performance, with particularly pronounced effects in multilingual settings where script variation, code-mixing, and sparse high-quality resources make conventional cleaning insufficient.

Table 16 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

5.3 Ablation Study: Safety and Toxicity Reduction

Beyond task-specific performance, our second ablation experiment investigates whether curation improves model safety, a critical consideration for deployment in sensitive linguistic and cultural contexts. We evaluate toxicity using the Toxigen benchmark via LLM360's Safety360 suite[7], which provides both explicit and subtle adversarial prompts spanning identity-based categories (race, religion, gender, nationality) and general offensive content. The evaluation protocol generates model completions from curated Toxigen templates, then classifies outputs using a RoBERTa-based [34] detector fine-tuned for nuanced and context-dependent toxic language detection. This methodology captures not only overt hate speech but also subtle stereotype amplification, coded language, and identity-based microaggressions that simpler keyword-based detectors miss. We compare our curated PARAM-1 [51] model against three multilingual baselines: SARVAM-1[8], LLaMA3.2-9T-3B [70, 71, 21], and Gemma2-2T-2B [64, 65, 66], all of which represent state-of-the-art multilingual language models trained on substantial corpora but without our specialized pipeline.

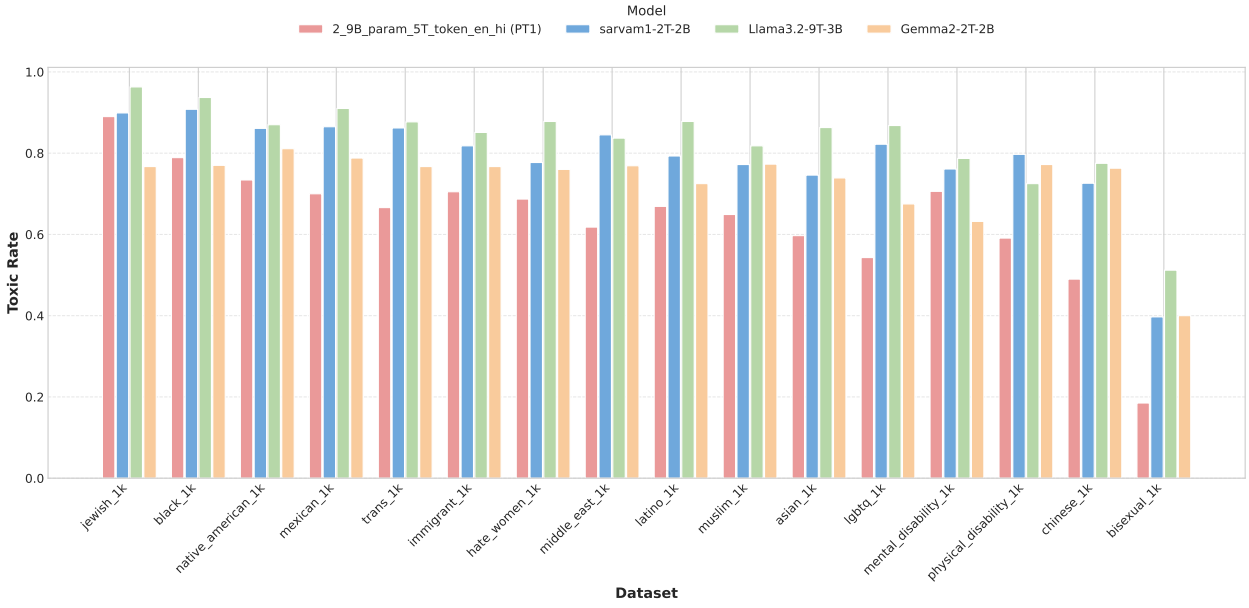


Figure 6 Toxicity Comparison

[7] https://github.com/LLM360/Analysis360
[8] https://www.sarvam.ai/

Figure 6 visualizes toxicity rates across 16 datasets encompassing both neutral baseline prompts and adversarial examples designed to elicit harmful outputs. The curated PARAM-1 model consistently maintains lower toxicity rates than all three baselines across nearly every evaluation condition. On adversarial identity-based prompts, the most challenging category where models are explicitly prompted to generate stereotypical or prejudiced content,PARAM-1 demonstrates particularly strong resistance, producing toxic outputs at rates 15-30% lower than comparable models. Critically, this improved safety profile does not come at the cost of over-censorship or reduced utility: the model continues to generate fluent, contextually appropriate responses to neutral prompts while declining to amplify harmful stereotypes or engage with bad-faith adversarial framing. These results establish that our two-stage toxic filtering process, combining rule-based initial flagging with multilingual RoBERTa-based reclassification, effectively removes training examples that would otherwise teach models to reproduce harmful patterns, without introducing excessive false positives that would degrade linguistic coverage or cultural representativeness.

Taken together, these ablation experiments provide robust evidence that our curation pipeline delivers dual benefits: substantial improvements in task-specific accuracy and reasoning capabilities, alongside meaningfully safer generation behavior that avoids stereotype amplification and identity-based harm. The performance gains on Hindi benchmarks and the reduced toxicity rates in multilingual contexts are particularly noteworthy, demonstrating that careful attention to data quality yields compounding returns for low-resource languages where existing models struggle most. By combining language-specific quality classification, synthetic rewriting for medium-quality content, aggressive deduplication, nuanced toxic content filtering, and comprehensive PII redaction, we establish a strong foundation for responsible deployment of large language models in multilingual and culturally diverse settings. This curation infrastructure not only enhances MILA's capabilities but provides reusable patterns for future corpus construction efforts aimed at equitable language technology development across the world's linguistic diversity.

6 OCR Processing: Digitizing Indic Scripts at Scale

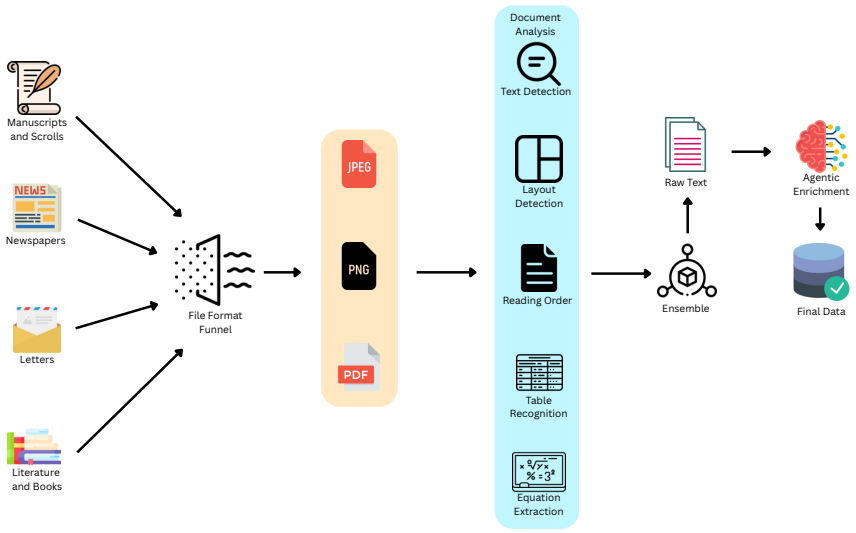


Figure 7 OCR Processing

OCR constitutes a critical bottleneck in constructing large-scale Indic language corpora, directly addressing the fundamental scarcity of digitized content across India's linguistic landscape. While high-resource languages like Hindi and Tamil possess partial digital presence through government digitization initiatives and online publishing, low-resource languages such as Maithili, Sindhi, Konkani, and Bodo remain largely trapped in print form: textbooks, literature, government documents, and historical archives that exist only as physical artifacts. To bridge this digital divide, we collected and processed 5-6 million pages from print materials and scanned books, converting otherwise inaccessible knowledge into machine-readable form suitable for language model training. This OCR pipeline represents not merely a technical preprocessing step but a fundamental act of linguistic preservation and democratization, enabling computational access to cultural and educational resources that have historically been confined to physical libraries and institutional archives.

6.1 Technical Challenges in Indic Script Recognition

The technical challenges inherent in Indic OCR substantially exceed those encountered in Latin-script languages. Approximately 37 percent of our scanned corpus exhibited severe degradation: faded ink from aging paper, irregular printing quality from legacy presses, and evolving orthographies that complicate character recognition across historical documents. These artifacts necessitated extensive preprocessing involving denoising filters, contrast enhancement algorithms, and binarization techniques to extract legible inputs from deteriorated source material. Yet even high-quality modern scans posed fundamental recognition challenges rooted in the structural complexity of Indic scripts themselves. Unlike alphabetic writing systems where characters map cleanly to phonemes, Indic alphabets employ stacked ligatures where consonant clusters merge into composite glyphs, conjunct consonants that span multiple Unicode codepoints, and extensive use of diacritics that modify base characters in visually intricate ways. Generic OCR systems trained predominantly on Latin scripts fail catastrophically on these features, producing broken Unicode sequences, incorrect ligature decompositions, and spacing artifacts that render output text unusable for downstream language modeling. The diversity of scripts: Devanagari, Bengali, Tamil, Telugu, Gujarati, Kannada, Malayalam, Odia, Gurmukhi, and others; further compounds the challenge, as each script exhibits unique glyph formation rules, different rendering conventions, and distinct sets of conjuncts that demand specialized recognition models.

Beyond script-level complexity, document layout heterogeneity introduced additional failure modes. Our corpus spans textbooks with multi-column layouts and embedded diagrams, newspapers with dense text interspersed with advertisements, manuscripts with marginalia and annotations, and technical documents containing tables and mathematical expressions. Naively applying off-the-shelf OCR to such varied material produced garbled output where column boundaries were ignored, reading order was scrambled, and non-textual elements were misinterpreted as corrupted text. This demanded tailored pipelines, as shown in Figure 7, incorporating layout analysis algorithms that segment pages into logical regions, line detection systems aware of script-specific baselines, and text normalization procedures that repair common OCR errors like spurious spaces within words or merged tokens across line breaks. To improve fidelity, we implemented confidence-based routing where OCR outputs were stratified by per-character confidence scores, high-confidence text proceeded directly to the corpus while low-confidence segments triggered language-specific correction layers. These correction modules leveraged contextual language models fine-tuned on clean Indic text to repair ligature breaks, restore dropped diacritics, and reconstruct plausible Unicode sequences from corrupted output. This iterative refinement process highlighted a fundamental interdependence: robust OCR requires curated training data to learn script-specific patterns, yet high-quality data creation depends on reliable OCR pipelines to digitize source material at scale. Breaking this circular dependency required simultaneous investment in both OCR model development and manual annotation efforts to bootstrap the system.

6.2 ISOB-Small: A Synthetic Benchmark for Indic OCR

To rigorously evaluate OCR quality and guide pipeline improvements, we developed the Indic Synthetic OCR Benchmark (ISOB-Small), a controlled testbed spanning 22 Indian languages across 110 synthetically generated pages. Direct evaluation on real scanned documents proved infeasible due to copyright restrictions on the archival materials obtained through formal memoranda of understanding with institutional partners. Rather than compromising evaluation rigor, we designed ISOB-Small to systematically reproduce the challenges observed in real-world digitization through synthetic generation. Each page in ISOB-Small incorporates realistic degradations encountered during document processing: multi-column layouts that test layout analysis, dense tables and figures that challenge region segmentation, mathematical expressions with mixed scripts, watermarks and stamps that introduce visual noise, paper folds and shadows that create uneven illumination, font variations that stress glyph recognition, and controlled blur levels that simulate aging or poor scan quality. By programmatically generating these artifacts from clean ground truth text, ISOB-Small provides a copyright-compliant evaluation framework while exposing the specific failure modes that generic OCR systems exhibit on Indic scripts. Representative examples from the benchmark are shown in Figure 8, illustrating the diversity of layouts, scripts, and degradations included in the test set.

The benchmark creation pipeline begins with seed corpus selection from existing OCR'd pages in hOCR format, filters for high-difficulty documents using confidence scores and VLM-based complexity prediction, randomly selects 3-10 target languages, extracts a taxonomy of hard artifacts from real documents, augments the selected pages with these artifacts while translating text into target languages, renders the enriched hOCR into visual form, applies style transformations using image editing models prompted with Indian manuscript aesthetics, and finally introduces low-level image augmentations including orientation changes, contrast shifts, noise injection, and geometric distortions. This

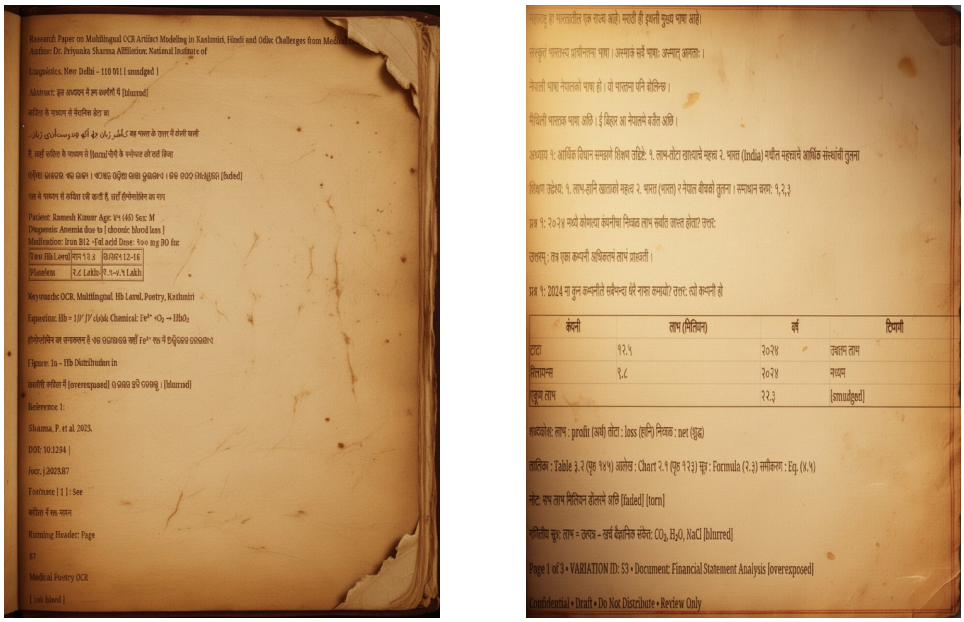


Figure 8 Samples from ISOB Benchmark

systematic generation process ensures comprehensive coverage of OCR challenges while remaining fully reproducible and extensible to additional languages or artifact types.

Beyond immediate corpus construction, ISOB-Small represents a foundational contribution to the broader research community working on low-resource language digitization. Recognizing that legal and ethical constraints prevent many researchers from accessing real archival materials for evaluation, we are releasing not only the benchmark dataset itself but also the complete generative recipes and code infrastructure used to construct it. This enables researchers to extend ISOB-Small to additional languages, generate larger test sets with customized difficulty profiles, or create domain-specific variants targeting particular document types such as legal texts, scientific literature, or historical manuscripts. The synthetic generation approach circumvents copyright barriers while providing controlled evaluation that isolates specific OCR challenges, a methodology applicable far beyond the Indic language context to any script or language lacking adequate digitization benchmarks. By open-sourcing both data and methodology, we aim to accelerate progress on OCR for underserved writing systems worldwide, fostering reproducible experimentation and enabling fair comparison across OCR systems and postprocessing techniques.

6.3 Comparative Evaluation and Postprocessing Impact

Evaluation on ISOB-Small revealed stark performance disparities across OCR systems and exposed critical weaknesses in vision-language models applied to Indic text recognition. Table 17 presents comprehensive results on both existing Indian OCR benchmarks (Bhashini) and the synthetic ISOB-Small testbed, tracking character error rate (CER), word error rate (WER), position-independent word error rate (PI-WER), and character-level 3-gram F1 scores. Specialized OCR models such as DotsOCR and Surya achieve substantially lower error rates on Bhashini (CER of 0.168 and 0.2 respectively) compared to general-purpose vision-language models like Qwen2.5-VL-72B [2, 77, 52, 15] (CER of 0.676) and Llama-4-Scout [70, 71, 21] (CER of 0.259). While VLMs offer broader task coverage and can handle diverse document types without specialized training, they suffer from hallucination on Indic scripts—generating plausible-looking but semantically incorrect text that is difficult to detect through automated metrics alone. In contrast, traditional OCR models produce more predictable error patterns that can be systematically addressed through postprocessing rules and language model correction. The performance gap widens dramatically on ISOB-Small, where VLMs struggle with the synthetic complexity: models like pixtral-12B [1] and GLM-4.1V-9B-Thinking [81] exhibit CER exceeding 4.0, while SmolDocling [40] and InternVL [6] variants fail catastrophically with error rates above 38. These results validate ISOB-Small as a genuine stress test that exposes brittleness in systems that perform adequately on cleaner benchmarks.

Table 17 Model Performance Benchmarks for Different Pipelines

List of Models	Bhashini				Mozhi			
	CER	WER	PI-WER	Char3 F1	CER	WER	PI-WER	Char3 F1
DotsOCR	0.168	0.253	0.23	0.801	0.12	0.19	0.9	0.88
Surya	0.2	0.28	0.138	0.867	0.14	0.21	0.91	0.89
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	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
SmolDocling-256M-preview	1.235	1.4	0.988	0.016	137.66	55.57	0.946	0.0001
RolmOCR	1.938	2.019	0.498	0.552	986.91	263.08	0.692	0.111
olmOCR-7B-0825	2.068	1.842	0.516	0.531	28.53	6.48	0.704	0.126
Nanonets-OCR-s	3.573	2.318	0.568	0.471	305.27	42.57	0.685	0.161
GLM-4.1V-9B-Thinking	4.384	3.88	0.893	0.08	755.1	321.92	0.985	0.0001
MinerU2.5-2509-1.2B	5.176	3.214	0.906	0.095	180	55	0.91	0.1
pixtral-12B	5.847	4.86	0.893	0.163	1.47	0.999	0.941	0.0039
InternVL3_5-GPT-OSS-20B-A4B-Preview-HF	38.87	4.537	0.994	0.0029	195.85	240.2	0.919	0

The critical importance of postprocessing becomes evident when examining performance improvements after applying our language-specialized correction pipeline. Table 18 demonstrates that targeted enhancements—including dictionary-based correction, language model rescoring, ligature repair, and Unicode normalization—substantially reduce error rates even for already-strong baselines. DotsOCR improves from 0.168 to 0.085 CER on Bhashini after postcorrection, while Surya advances from 0.2 to 0.095 CER. These gains translate directly to corpus quality: reducing CER by half means doubling the amount of usable training data extracted from each scanned page. On ISOB-Small, postcorrected models achieve scores above 0.86, confirming that the combination of specialized OCR with targeted postprocessing provides a robust solution for Indic digitization. These targeted enhancements proved critical not only for benchmark performance but for producing high-quality, machine-readable text suitable for downstream language model training.

Table 18 Model Performance on Benchmarks

Preprocessing / Post-Correction Performance Bhashini

List of Models	CER	WER	PI-WER	Char3 F1
Dots.OCR - postcorrected	0.085	0.145	0.12	0.91
Surya - postcorrected	0.095	0.16	0.11	0.925

ISOB-Small Results

List of Models	ISOB-Small
Dots.OCR	0.8616
Surya	0.8982

6.4 LLM-Assisted Quality Evaluation

To validate that improvements in traditional metrics (CER, WER) actually translate to better semantic quality, we designed a comprehensive evaluation framework leveraging large language models as quality judges. Conventional OCR metrics measure surface-level string similarity but fail to capture whether postprocessing interventions—such as sentence reordering for improved coherence, ligature repair that changes character sequences, or contextual corrections that substitute semantically equivalent terms—preserve or enhance meaning. This limitation is particularly acute for quality-enhanced text where deliberate modifications may increase string distance from raw OCR while improving linguistic fidelity. Our LLM-assisted evaluation addresses this gap through multi-stage assessment using state-of-the-art models including GPT-OSS-120B [43], Deepseek [11], and Qwen [2, 77, 52, 15]. Each model is prompted to compare original ground truth with both raw OCR and postprocessed outputs, providing consistency scores that reflect semantic preservation independent of surface form. We complement direct text comparison with multilingual embedding similarity metrics that capture semantic alignment in native language space, cross-lingual evaluation via back-translation where both ground truth and enhanced text are translated to English and compared using independent LLM judges, and vision-language similarity using CLIP and SIGLIP embeddings to measure structural fidelity between page reconstructions. Finally, we incorporate standard translation metrics including BLEU, ROUGE, and CHRF++ to provide quantitative baselines alongside the LLM judgments. This ensemble approach ensures robust quality assessment that captures improvements invisible to traditional metrics while avoiding over-reliance on any single evaluation paradigm.

6.5 Ablation Study: Conventional vs Processed OCR Data

The practical impact of OCR quality on downstream model training is demonstrated through controlled experiments on a 310M parameter dense language model. We compared training dynamics on two versions of the same corpus: raw OCR output with typical error rates around 15-20 percent, and quality-enhanced text after applying our full post-processing pipeline. Pretraining on raw OCR data produced highly unstable loss curves with frequent spikes, irregular perplexity behavior, and slow convergence, clear indicators of training corpus noise overwhelming the learning signal. Models trained on this noisy data struggled to learn coherent linguistic patterns, exhibiting high validation perplexity and poor performance on downstream tasks. In stark contrast, training on postcorrected text yielded smooth, monotonic loss reduction and stable perplexity curves characteristic of high-quality pretraining data, as visualized in Figure 16.

Architecture attributes	Values
Model Architecture	causal language model
Hidden size	768
Intermediate size (FFN)	3108
Max Position Embeddings	2048
Num of Attention Heads	12
Num of Hidden Layers	12
Num of Query Groups	12
Normalization	RMSNorm
Activation Function	swiglu
Attention Type	Multi-head Attention
Position Embedding Type	RoPE (rotary)
Dropout (hidden/attn/ffn)	0.0 / 0.0 / 0.0
Precision	bf16 (AMP O2)

Table 19 Architecture Details of 310M Parameter Model

The quality-enhanced model achieved substantially lower final perplexity and demonstrated stronger performance on Indic language benchmarks, confirming that preprocessing investments directly improve model capabilities. These results underscore a critical insight: for low-resource languages where data scarcity already constrains model quality, training on noisy OCR outputs compounds the disadvantage by forcing models to learn from corrupted signal. Conversely, investing in robust OCR pipelines and postprocessing infrastructure amplifies the value of scarce digitized resources, enabling smaller corpora to support stronger models.

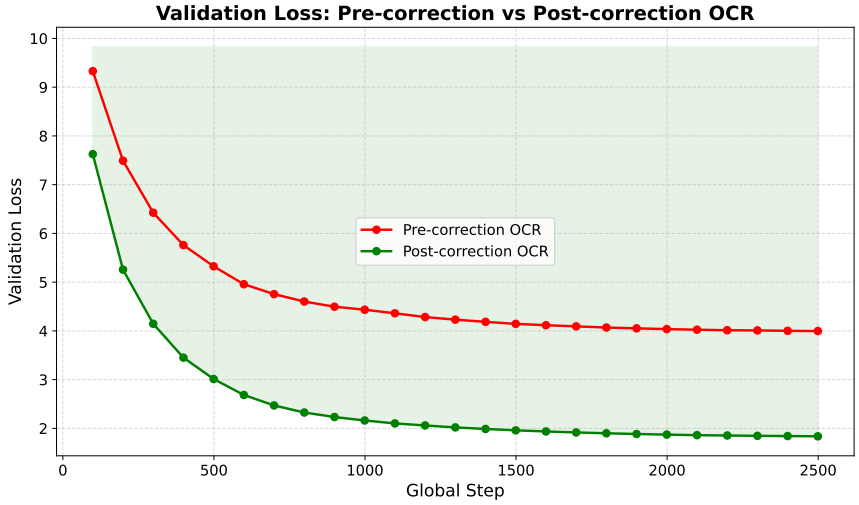


Figure 9 Validation loss comparison between models trained on conventional OCR output versus postcorrected text. The postcorrected corpus produces substantially more stable training dynamics and lower final perplexity, demonstrating the direct impact of OCR quality on model performance.

The infrastructure and validation framework established through this work thus extends beyond MILA itself, providing reusable tools that support equitable language technology development across the world's linguistic diversity.

7 Translation Pipeline

A fundamental challenge in constructing large-scale Indic datasets lies not merely in the absence of digital text, but in the structural scarcity of high-quality monolingual and parallel corpora that could enable cross-lingual transfer from resource-rich languages. Translation-based data generation directly addresses this bottleneck by producing parallel corpora that facilitate knowledge transfer from languages with abundant digital resources, primarily English, into the 16 Indic languages that form the backbone of MILA. This capability proves critical for multilingual model development, as recent work has demonstrated that downstream performance on mathematical reasoning, STEM domains, and code generation benefits substantially from exposure to parallel corpora during pretraining [5, 32]. The mechanism underlying these gains appears to be the model's ability to align semantic representations across languages, learning that mathematical operations, logical reasoning patterns, and algorithmic structures transcend linguistic boundaries. By providing models with paired examples that express identical concepts in multiple languages, parallel corpora enable the extraction of language-invariant knowledge structures that generalize more robustly than monolingual training alone.

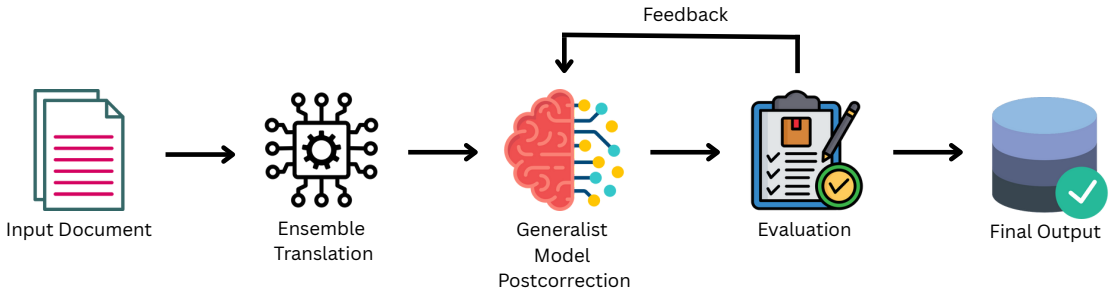


Figure 10 Translation Pipeline

7.1 The Specialist-Generalist Tension in Low-Resource Translation

The production of high-quality translations for low-resource Indic languages presents challenges that extend far beyond simply applying off-the-shelf systems. As demonstrated in Table 20, no single model achieves universally optimal performance across the 15 Indic languages we target. Dedicated machine translation systems such as IndicTrans2 (IT2) [16] perform strongly across a broad range of languages, particularly when combined with preprocessing and postprocessing pipelines, where they achieve consistent improvements over the raw IT2 outputs (e.g., +2–3 chrF++ on average). Nevertheless, IT2 still struggles with truly low-resource languages such as Maithili and Sanskrit, where limited parallel training data constrains model effectiveness (27.70–30.24 chrF++ on Sanskrit). By contrast, general-purpose multilingual models such as NLLB-200-3.3B [67], NLLB-moe-54B [67], DeepSeek V3.1 Think [11], and Llama-4-Maverick-17B [70, 71, 21] provide broader linguistic coverage without explicit fine-tuning. These generalist systems achieve competitive scores on well-resourced languages (e.g., NLLB-moe-54B reaches 57.03 chrF++ on Kannada; Llama-4-Maverick 57.34 on Telugu), but their performance remains uneven on low-resource settings, with noticeable drops on Sanskrit and Maithili. Moreover, their outputs often exhibit subtle semantic drift, morphological inconsistencies, or code-mixing with English or Hindi, which undermines corpus quality. Alternative approaches such as Hunyuan-MT [83] demonstrate catastrophic failure modes, producing near-zero scores on Nepali, Oriya, and Punjabi, underscoring the fragility of systems lacking robust multilingual grounding.

This fundamental tension between specialist precision and generalist coverage motivates our carefully architectured translation pipeline, as shown in Figure 10, which eschews reliance on any single system in favor of an ensemble approach that combines specialist and generalist models through multi-stage processing. Our pipeline begins with synthetic augmentation, translating content from English and other resource-rich languages into the 16 target Indic languages using an ensemble of specialist machine translation models and generalist large language models. Rather than treating translation as a one-pass operation, we implement a robust LLM-based post-correction phase that repairs syn-

tactic and semantic inconsistencies introduced during initial translation, enhances context preservation across sentence boundaries, and addresses the subtle morphological and syntactic variations that distinguish natural Indic language use from mechanically translated text. This post-correction stage leverages state-of-the-art multilingual language models, specifically those demonstrating strong performance on benchmarks as shown in Table 20 (b), to ensure that postprocessing genuinely improves linguistic quality rather than introducing new errors through models with insufficient Indic language proficiency.

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 20 Translation Model Performance on FLORES Benchmarks using chrF++

7.2 LLM-Based Post-Correction and Human Validation

The impact of this post-correction architecture becomes evident in Table 20 (b), where our processed IndicTrans2 (IT2) pipeline achieves substantial improvements over baseline translation quality. Post-correction elevates performance from 45.10 to 48.40 chrF++ for Assamese, from 48.67 to 51.71 for Bengali, and from 52.23 to 55.99 for Malayalam, demonstrating consistent gains across the entire language spectrum. More critically, the pipeline shows its greatest impact precisely on the low-resource languages where initial translation quality is weakest: Sanskrit improves from 27.70 to 30.24 chrF++, while Sindhi advances from 53.03 to 56.86. These gains translate directly to corpus utility, improving translation quality by 5-7 percent effectively increases the amount of usable training data by similar margins, as higher-fidelity translations reduce the noise that would otherwise corrupt the learning signal during language model pretraining. The consistent improvements across all 15 languages validate our hypothesis that ensemble translation followed by targeted LLM-based correction provides more robust quality than relying on any single translation system, however sophisticated.

Human evaluation plays a critical role at multiple stages of this pipeline, providing ground truth validation that guides both model selection and quality assessment. We integrated three expert language evaluators for each target language, who reviewed initial translation outputs to identify systematic error patterns, evaluate cultural appropriateness of linguistic choices, and ensure that translations reflect natural language use rather than the stilted, overly literal renderings characteristic of naive machine translation. This human-in-the-loop approach proved particularly valuable for detecting subtle issues invisible to automated metrics: gender agreement errors in morphologically rich languages like Hindi and Bengali, inappropriate register choices that clash with the formality level of source content, and culturally insensitive translations that preserve denotative meaning while losing connotative appropriateness. The evaluators' feedback directly informed our model selection process, leading us to preferentially weight outputs from models that demonstrated stronger alignment with natural Indic language patterns as judged by native speakers. This validation framework ensures that our pipeline produces not merely linguistically correct translations, but culturally aligned representations that will support language models in learning authentic Indic language use rather than absorbing artifacts of mechanical translation. We have showcased a study on readability of translation experiments in Section 10.

For the post-correction phase, we first identify the most suitable large language models (LLMs) on a per-language basis by consulting the Indic MMLU benchmark results (Table 3). This ensures that the models chosen for grammatical refinement are not only capable in general reasoning but also demonstrate strong competence in the specific Indic language of interest. For instance, DeepSeek V3.1 [11] was selected for Hindi owing to its consistently strong Indic MMLU scores, while Gemma-3 27B [64, 65, 66] was preferred for Tamil due to its comparatively better alignment on Dravidian languages. Once selected, these models are guided through carefully designed prompts that frame post-correction as an expert linguistic transformation task, positioning the model as a specialist in the target language with deep understanding of grammar, syntax, and natural conventions. The prompt explicitly directs the model to convert

poorly structured, grammatically incorrect, or awkward translations into natural, well-formed text while adhering to several critical constraints: strict semantic preservation with no omissions, exclusive use of the target language with no English or Hindi code-mixing, grammatical accuracy with natural flow, and complete avoidance of meta-commentary that would contaminate the corpus. This design reflects lessons learned from extensive experimentation: early prompt versions that omitted explicit prohibitions against code-mixing often produced outputs interspersed with English terms, while lack of explicit length guidance led to truncated or verbose outputs that distorted information density. The refined prompt achieves a careful balance, enabling models to repair genuine errors while preserving the integrity and content of the original translation.

Prompt Template for Post-Correction

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

The qualitative impact of this post-correction process becomes visible in Figure 11, which presents examples of text before and after processing. The improvements span multiple dimensions of linguistic quality: grammatical structures are regularized to match target language conventions, sentence flow is enhanced through appropriate use of discourse markers and transitional phrases, and awkward word choices are replaced with more natural alternatives that better capture the intended meaning. Critically, these transformations preserve semantic fidelity, the corrected text expresses the same propositional content as the original translation while rendering it in more fluent, idiomatic form. This balance between correction and preservation proves essential for corpus quality: overly conservative post-correction leaves errors intact, while overly aggressive rewriting risks introducing semantic drift that corrupts the training signal. Our prompt-guided approach navigates this tension by giving models clear objectives (improve naturalness, correct grammar) alongside explicit constraints (preserve meaning, maintain length), enabling reliable quality enhancement without the semantic instability that plagued earlier correction attempts.

7.3 Hierarchical Chunking for Long-Context Translation

Beyond improving the quality of individual translations, our pipeline addresses a fundamental challenge that emerges when translating long, complex documents: the context window limitations of language models and the difficulty of maintaining semantic coherence across extended sequences. This challenge proves particularly acute for technical content in mathematics, STEM domains, formal proofs, and code, where dependencies often span thousands of tokens and losing context mid-translation can corrupt the logical structure that makes such content valuable. To address this, we implement a hierarchical chunking strategy that divides long sequences into manageable segments while preserving cross-segment dependencies through carefully designed context propagation. Documents are segmented based on both token count constraints and logical unit boundaries, ensuring that each chunk remains within the model's context window while respecting natural breakpoints such as section boundaries in technical documents or function definitions in code. Each segment is provided with summaries of preceding chunks, enabling the model to maintain continuity of reasoning and avoid the abrupt semantic discontinuities that arise when translating isolated fragments. For highly structured content such as multi-step mathematical proofs or interdependent code blocks, we introduce overlapping tokens between consecutive chunks, creating redundancy that prevents loss of critical context at segment boundaries.

This chunking strategy enables coherent translation of documents that would be intractable under single-pass approaches, but introduces new challenges around ensuring consistency across segment boundaries. A mathematical proof translated in three separate chunks might use different terminology for the same concept, or a code snippet might employ inconsistent naming conventions because the model lacks awareness of choices made in earlier segments. To

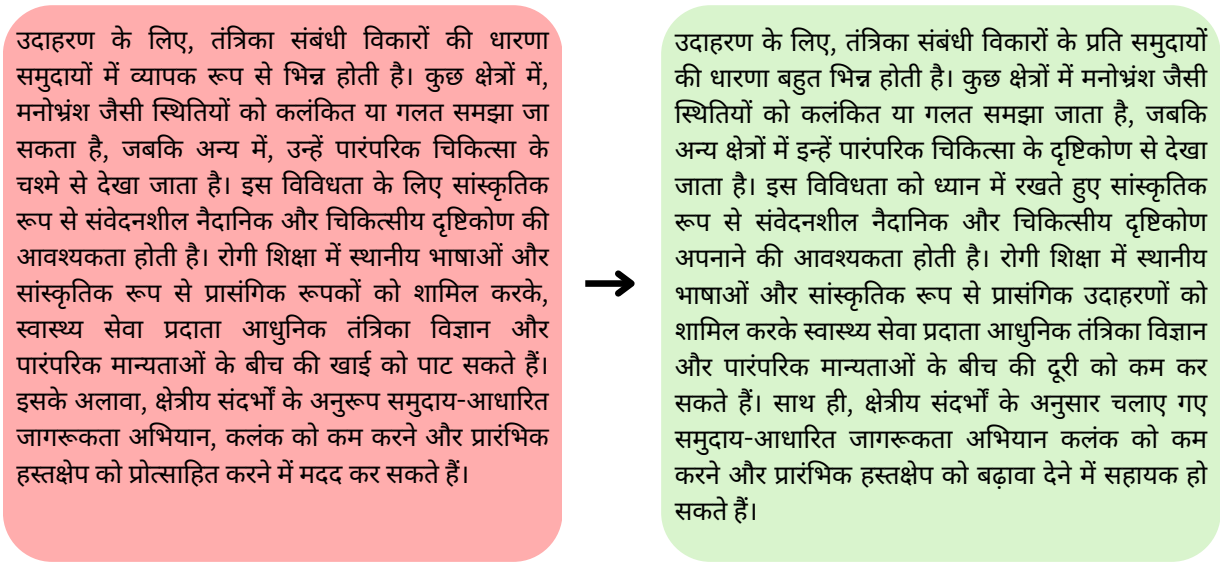


Figure 11 Example of translation post-correction showing improvements in grammatical structure, sentence flow, and natural language use while preserving semantic content. The corrected text demonstrates enhanced fluency and cultural appropriateness compared to raw machine translation output.

address these consistency challenges, we implement a multi-stage validation process that operates after initial chunked translation is complete. Consecutive translated chunks are evaluated using large language models as coherence judges, with prompts designed to detect terminology drift, verify continuity of meaning, and identify inconsistencies in technical notation or variable naming. This coherence validation stage flags problematic boundaries for human review, enabling linguists to harmonize terminology and ensure that the final concatenated document reads as a unified whole rather than a patchwork of independently translated fragments. We complement this LLM-based coherence checking with embedding similarity metrics that quantify semantic alignment across chunk boundaries, providing quantitative signals that supplement the qualitative judgments of language models and human reviewers, similar to those shown in Tables 1 and 2.

7.4 Downstream Impact and Infrastructure Integration

The complete translation pipeline thus operates as an ensemble system combining multiple complementary approaches: specialist machine translation models provide initial high-quality translations for well-resourced language pairs, generalist large language models offer broader coverage for low-resource languages and handle diverse content types, LLM-based post-correction repairs grammatical and stylistic issues while preserving semantic fidelity, hierarchical chunking enables translation of long technical documents that exceed single-pass context limits, coherence validation ensures consistency across segment boundaries, and human evaluation provides ground truth quality assessment that guides model selection and validates final outputs. This multi-stage architecture reflects a fundamental insight: for low-resource languages where no single system achieves adequate quality, robust translation requires carefully orchestrated combination of multiple imperfect systems, with each stage designed to address specific failure modes observed in preceding stages. The result is a translation pipeline that generated coherent, high-fidelity translations across billions of tokens, significantly reducing the errors and inconsistencies that would arise from naive single-pass translation approaches.

The impact of this translation infrastructure extends beyond immediate corpus construction to enable several downstream capabilities critical for MILA's overall architecture. First, the production of high-quality parallel corpora allows us to leverage cross-lingual transfer for improving model performance on truly low-resource languages where monolingual data remains insufficient even after extensive crawling and OCR efforts. By training models on aligned English-Indic parallel data, we enable them to bootstrap linguistic knowledge from high-resource languages while learning language-specific morphological and syntactic patterns from the Indic side of parallel pairs. Second, translated content

serves as a foundation for synthetic data generation, providing seed examples that guide persona-based augmentation pipelines in producing culturally appropriate and linguistically natural synthetic training data. Third, the translation of benchmarks such as MMLU into Indic languages provides evaluation infrastructure that enables fair comparison of model capabilities across languages, addressing the evaluation gap that has historically obscured the true performance of multilingual models on low-resource languages. Fourth, back-translation of Indic-language outputs into English enables quality assessment through comparison with original source text, providing a complementary validation signal that supplements direct Indic-language evaluation. Together, these capabilities position the translation pipeline not as an isolated preprocessing stage but as a foundational infrastructure layer that supports multiple aspects of MILA's data curation, augmentation, and validation workflows, enabling the construction of a truly multilingual corpus that serves the diverse needs of Indic language model development.

8 Synthetic Rewriting and Data Distillation

Data distillation addresses a fundamental paradox in contemporary multilingual language model development: while large language models can generate vast quantities of text [9, 47], this content predominantly reflects Western perspectives, embedding English-centric knowledge and cultural norms that fail to capture nuances essential for Indian applications. Even well-intentioned translations of datasets such as Sangraha Synthetic [27] preserve linguistic surface forms while fundamentally missing cultural context and domain-specific knowledge frameworks that distinguish authentically Indian content from mechanically translated approximations. A medical reasoning dataset translated into Hindi might use correct grammar while discussing healthcare delivery models entirely irrelevant to India's public health infrastructure and Ayurvedic traditions. Legal reasoning examples from Western jurisprudence fail to capture India's parallel legal systems where civil law coexists with personal laws rooted in religious traditions.

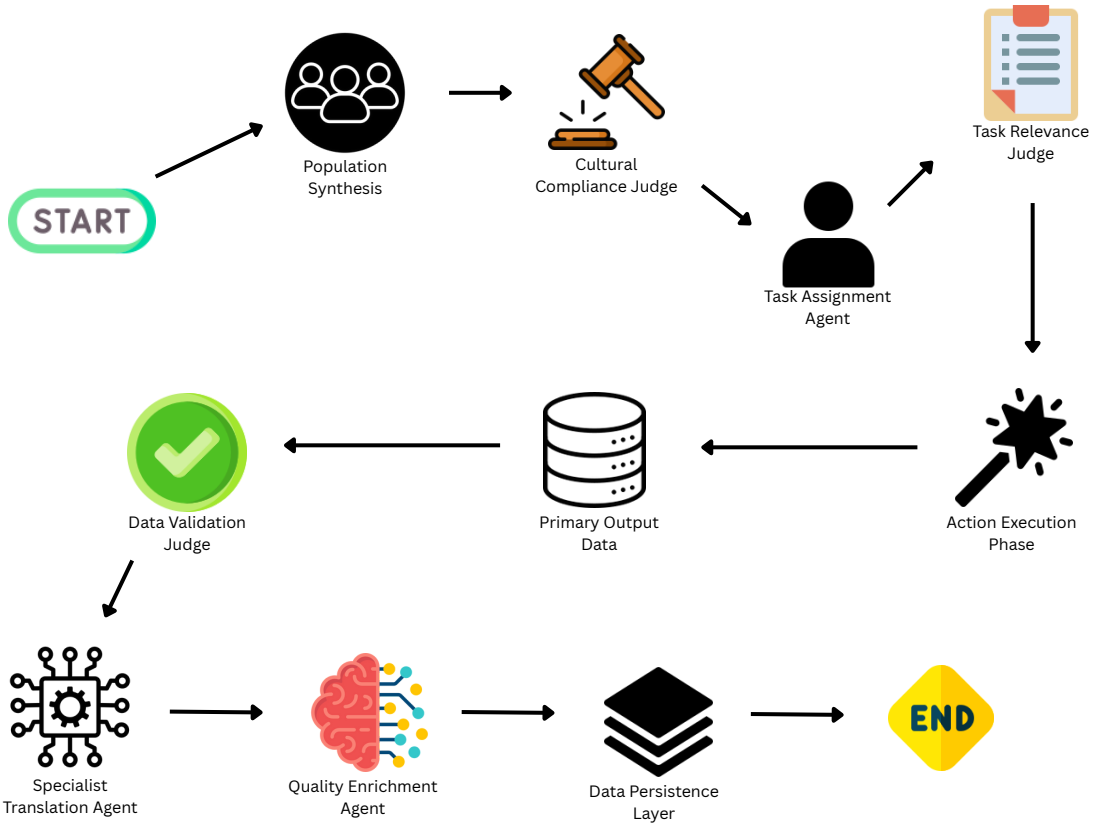


Figure 12 Multi-data distillation pipeline with iterative feedback loops ensuring rigorous quality control through cultural compliance, task assignment, relevance checking, validation, translation, and enrichment stages.

Distillation provides a solution by offering efficient access to world knowledge from state-of-the-art models while

enabling domain control and cultural shaping. Since strong language models specifically trained for Indian contexts remain unavailable, we leverage existing multilingual LLMs demonstrating high Indic-MMLU performance (Table 3) to systematically embed Indian values and reasoning patterns into synthetic generation. By distilling knowledge from powerful models while carefully controlling cultural framing through persona specification and generation constraints, we achieve both data efficiency and controllable alignment with local requirements.

8.1 Indic PersonaHub: Engineering Cultural Identity at Scale

To operationalize culturally grounded distillation, we developed Indic PersonaHub, a large-scale repository of over 300 million Indian personas spanning 1400+ domains. The motivation behind PersonaHub lies in a fundamental challenge: most foundation models, trained on predominantly Western-centric corpora, default to perspectives, reasoning patterns, and cultural assumptions that misalign with Indian discourse. Without intervention, even sophisticated downstream adaptations risk reproducing these biases, thereby diluting authenticity in tasks that require contextually grounded knowledge. PersonaHub addresses this gap by serving as a structured mechanism to embed Indian perspectives directly into synthetic training data.

Sample Persona

You are a patriotic Indian poultry farmer deeply committed to the nation's agricultural self-reliance and food security. They are vigilant about potential health risks and economic impacts of poultry diseases, particularly in the context of India's thriving backyard poultry sector. As a proud member of the Indian Poultry Farmers Association, they actively promote indigenous poultry breeds and sustainable farming practices. They are well-versed in disease management protocols and quarantine measures, recognizing their importance in safeguarding India's poultry industry. Their concerns extend to protecting commercial poultry farms, which are vital to India's rural economy and protein supply. They view their work as a service to the nation, aligning with the government's initiatives to boost agricultural productivity and rural livelihoods. 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 field of Poultry Farming and Livestock Management in India. 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: Poultry Farming and Livestock Management in India 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 be also 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 # How can India balance the growing demand for poultry and livestock products with the urgent need for sustainable and ethical farming practices, while addressing the socioeconomic challenges faced by small-scale farmers and ensuring food security for its vast population?

STYLE # Follow the simple writing style common in communications. Be persuasive yet maintain a neutral tone. Avoid sounding too much like 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.

Unlike demographic templates or shallow role labels, a persona in our framework is conceived as a multidimensional specification that guides language models toward authentic and culturally resonant responses. Inspired by billion-scale persona generation efforts [18], our design extends beyond age, gender, or occupation to encode rich contextual detail: linguistic preferences, cultural traditions, domain expertise, regional affiliations, and value systems. For example, an Ayurvedic practitioner from Kerala approaches medical reasoning through centuries, old holistic frameworks, while an

AIIMS, trained allopathic physician employs biomedical evidence and global clinical guidelines. Similarly, a Tamil literary scholar interprets texts through Sangam literature rather than Western critical theories. Such grounding ensures that personas operate not as abstract demographic placeholders but as situated voices representing diverse Indian knowledge systems.

A critical dimension of construction involves what we term Indianization of personas. This refers to the systematic transformation of generic or globally trained personas into forms that reflect Indian socio-cultural realities. Without explicit Indianization, a persona such as a "climate change scientist" defaults to discourses around carbon markets and individual consumption, topics common in Western debates but misaligned with India's climate discourse, which foregrounds equity in development, historical responsibility of industrialized nations, and tensions between growth and sustainability. To operationalize Indianization, each generated persona undergoes a cultural compliance review conducted by an LLM-based agent. This review evaluates whether the persona reflects appropriate cultural values, avoids stereotypes, and embodies authentic Indian perspectives rather than Western projections. Personas failing this evaluation are recycled through modified prompts until compliance is achieved. Validated personas are then assigned to functional tasks within relevant domains, ensuring both expertise alignment and cultural fidelity.

The example persona in this subsection highlights the granularity and cultural depth of PersonaHub. Rather than a neutral occupational profile, the poultry farmer is framed through national agricultural priorities, indigenous breeds, and socio-economic realities of small-scale farmers, perspectives that resonate strongly within Indian discourse but would be absent or flattened in generic datasets. By operationalizing Indianization at scale, PersonaHub establishes a foundation for culturally faithful synthetic data generation, enabling downstream models to reflect authentic Indian voices across diverse domains.

8.2 Culturally-Grounded Text Generation: From Persona to Production

The complete distillation pipeline (Figure 12) implements a multi-stage architecture designed to combine the scale of synthetic generation with rigorous quality assurance. At every stage, feedback loops and cultural checks ensure that only thoroughly vetted material enters the corpus. This reflects a central principle: scale alone is insufficient, and large volumes of synthetic data must be continuously filtered for coherence, factual reliability, and cultural resonance. The process begins with population synthesis, where more than 300 million raw personas are generated across 1400+ domains. While this stage establishes the breadth of identities, it also introduces the challenge of refinement, since not all personas are equally coherent or contextually appropriate. To address this, persona–task pairs are evaluated for relevance, verifying that each persona is logically aligned with the task it is assigned. A Carnatic musician persona, for instance, cannot meaningfully generate content for Hindustani gharana traditions, just as a Vedic mathematics expert approaches teaching with assumptions that diverge from those of a computational number theory specialist. Pairs that fail this relevance check are returned for reassignment rather than advanced to later stages, preventing the accumulation of weak alignments that would otherwise dilute the corpus.

Once coherence is ensured, approved pairs move to action execution, where persona-guided prompts generate the primary textual outputs. These outputs undergo multiple layers of validation, checking for factual accuracy, internal consistency, and cultural alignment with the persona's framing. Validated material then proceeds to translation, where specialist agents produce high-quality Indic language text. This stage leverages ensemble methods and post-correction techniques from our translation pipeline, ensuring outputs read as natural, idiomatic language rather than mechanically translated text carrying cross-lingual artifacts. A final enrichment stage polishes fluency and coherence, ensuring that the stored corpus balances the advantages of synthetic scale with the standards of linguistic and cultural quality required for downstream use.

The defining innovation of this pipeline lies in its integration of Indianization as a structural principle rather than a superficial add-on. Indianization is implemented in three interconnected layers that progressively embed cultural grounding into the generation process. Personas are first Indianized, transforming them from generic or globally influenced templates into culturally authentic actors situated within Indian contexts. Next, tasks are reframed through deeply reflective, domain-specific questions that stimulate reasoning grounded in Indian perspectives. A question such as "How should India navigate the ethical complexities of stem cell research, balancing medical advancement with moral considerations regarding embryonic cells, informed consent, and equitable access?" forces the model to reason within Indian ethical, social, and political frameworks rather than adopting universalized Western assumptions. Finally, the generation stage produces extended passages written in natural English but resonating with Indian audiences by incorporating cultural familiarity, narrative conventions, and societal norms.

By embedding these cultural interventions at every level, the pipeline transforms synthetic generation into a genuinely adaptive process. Translation can produce grammatically correct Indic text, but only culturally grounded synthesis generates content that reflects how Indians reason about complex issues and communicate within their discourse traditions. This layered design ensures that cultural authenticity is not treated as ornamentation but institutionalized as a non-negotiable dimension of quality, allowing the corpus to scale while remaining faithful to the perspectives it seeks to represent.

8.3 Structured Knowledge Extraction: QA and Instruction Dataset Construction

Complementing persona-driven synthesis, our QA extraction pipeline transforms unstructured Indic text into high-quality instruction data through four-stage processing. Context-aware chunking segments raw text into 1000-4000 token spans, preventing mid-sentence breaks and preserving logical coherence. Each segment is interpretable as standalone unit, critical for question generation where questions must be answerable from chunk information alone. Chunking respects document structure, treating section boundaries, paragraph breaks, and functional definitions as natural segmentation points maintaining semantic integrity.

Each chunk undergoes relevance checking and domain classification, filtering ephemeral or Western-centric content while assessing cultural relevance to Indian contexts. Valid segments receive domain assignments (Healthcare, Finance, History, Culture, BFSI, Education, Governance, Law, News, Sports, Tourism) through multi-label classification recognizing content often spans domains, India's pharmaceutical industry bridges Healthcare, Business, and Governance simultaneously. The pipeline processed 1121 chunks from Wikipedia Indic articles, 619 from DharmaWiki covering religious and philosophical traditions, and 4775 from diverse sources including government reports and news archives, yielding corpus balancing encyclopedic knowledge with culturally specific content underrepresented in Western knowledge bases.

From validated chunks, the pipeline generates fully self-contained questions spanning general explanation (comprehension), commonsense reasoning (implicit cultural knowledge), causal reasoning (exploring relationships), and open-ended prompts (inviting analysis). This diversity ensures models develop multiple reasoning capabilities beyond fact retrieval. Each question receives two answer forms: crisp answers for fact-based queries, and detailed answers (3-5 sentences) supplying explanatory context connecting questions to broader domain knowledge. This multi-fidelity generation recognizes different use cases: conversational agents benefit from detailed contextually rich responses while fact-checking systems require concise verifiable statements.

8.4 Ablation Experiment: Conventional vs Distilled Downstream Performance

The empirical impact of comprehensive distillation, combining persona-driven synthesis with structured QA extraction, is shown in Table 21, comparing models fine-tuned on open source data against our in-house recipe. Across thirteen diverse benchmarks spanning commonsense reasoning, knowledge assessment, and truthfulness, our distilled data demonstrates consistent improvements. HellaSwag [80] accuracy advances from 70.47 to 73.07, with Hindi version improving from 44.01 to 44.59. More dramatic gains appear on challenging tasks: MMLU Pro [73] improves from 5.23 to 8.73 exact match (67% relative improvement), while CommitmentBank [10] advances from 30.36 to 57.14, nearly doubling performance. Indic-specific evaluations show MILU [72] Hindi improving from 28.87 to 32.26 and Sanskriti States from 55.13 to 55.91. Consistency across diverse task types validates our hypothesis that culturally grounded, persona-driven synthetic data provides training signal qualitatively different from mechanically curated or translated datasets.

By combining synthetic article generation anchored in culturally grounded personas with structured QA extraction from authentic Indic sources, the distillation pipeline achieves both breadth and depth. The synthetic component provides scale and domain coverage impossible through manual curation alone, while extraction ensures grounding in real-world Indian knowledge sources and linguistic patterns. The result is a resource that is factually grounded, instruction-ready, and culturally resonant—uniquely suited for fine-tuning language models for Indian contexts. This infrastructure represents not merely a data processing pipeline but a fundamental rethinking of training data construction for low-resource, culturally distinct linguistic communities in an era dominated by English-centric foundation models trained primarily on Western corpora.

Table 21 Comparison of Open Source SFT and In-house SFT Data Recipe across different tasks.

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	37.40
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	37.70
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

9 Data Organisation

Building a multilingual Indic dataset spanning 7.5 trillion tokens presents governance challenges that fundamentally differ from those encountered in conventional English corpus construction, as emphasized by foundational work on dataset documentation and transparency [19, 23]. Unlike English corpora, which benefit from decades of standardization efforts, established digitization practices, and relatively uniform encoding conventions, Indic data confronts structural fragmentation across multiple dimensions simultaneously. Sources span digitized textbooks with varying OCR quality, newspapers employing inconsistent orthographic conventions, government documents using legacy encoding schemes, social media content mixing scripts and languages within single posts, and historical archives where material has been digitized under different technical standards across decades. This heterogeneity manifests not merely as noise to be filtered but as fundamental diversity requiring preservation—the very linguistic variation that makes low-resource languages distinct risks being erased through overly aggressive normalization. The challenge extends beyond scale to encompass control: ensuring that truly low-resource languages are preserved in the long tail of the distribution rather than being overwhelmed by higher-resource languages, verifying that Unicode normalization operations do not inadvertently collapse phonologically distinct characters that appear visually similar across scripts, and maintaining complete auditability of every transformation applied to source material so that downstream model behaviors can be traced back to specific data processing decisions.

Without rigorous governance infrastructure and systematic taxonomic organization, a trillion-token Indic dataset risks becoming simultaneously too brittle for production deployment and too opaque for scientific reproducibility. Brittleness emerges when licensing restrictions are inadequately tracked, causing models trained on the corpus to inherit legal liabilities; when personally identifiable information leaks through inadequate filtering; or when quality degradation in specific language-domain combinations goes undetected because monitoring lacks the granularity to surface issues affecting small subpopulations. Opacity arises when transformation lineage is lost, making it impossible to debug model behaviors by examining training data provenance; when versioning is ad-hoc, preventing reproducible experimentation; or when metadata is incomplete, leaving researchers unable to construct domain-specific subsets or balance corpus composition across linguistic and topical dimensions. These governance failures compound in low-resource language contexts, where the community lacks the scale to absorb quality issues through redundancy and where each dataset artifact represents irreplaceable cultural and linguistic resources that cannot be easily regenerated if corrupted or lost.

9.1 Lakehouse Architecture: Unifying Storage, Metadata, and Governance

To address complex governance requirements, we implement a petabyte-scale AI data lakehouse architecture that unifies storage, lineage tracking, metadata cataloging, governance enforcement, and versioning. Unlike traditional data lakes, which offer scale without governance, or warehouses, which enforce governance but lack flexibility, the lakehouse paradigm combines their strengths, supporting ACID transactions and schema enforcement alongside schema-on-read adaptability for diverse machine learning corpora. The storage layer, built on JuiceFS with MinIO object

storage, is organized into three zones reflecting data maturity. The Raw zone ingests source material in original form, preserving even malformed or duplicate documents to safeguard scarce Indic resources. The Curated zone stores cleaned, deduplicated, and standardized data with full metadata, while the Feature Store holds processed features such as tokenized sequences and embeddings, versioned with their generating code for reproducibility. These tiers enforce governance checkpoints and allow safe experimentation without contaminating production data.

Lineage tracking employs OpenLineage and Marquez to capture transformation events across heterogeneous tools such as Spark, Airflow, and Kafka. OpenLineage provides a vendor-neutral specification, while Marquez aggregates lineage into a queryable graph. This enables forward queries (which artifacts depend on this source?) and backward queries (which sources produced this feature?), supporting debugging, provenance verification, and compliance. When a model exhibits anomalies on specific Indic examples, lineage reveals the complete processing path, feature engineering, cleaning, translation, or OCR, back to the source document. It also ensures auditors can verify that sensitive or licensed content is correctly handled across pipeline stages.

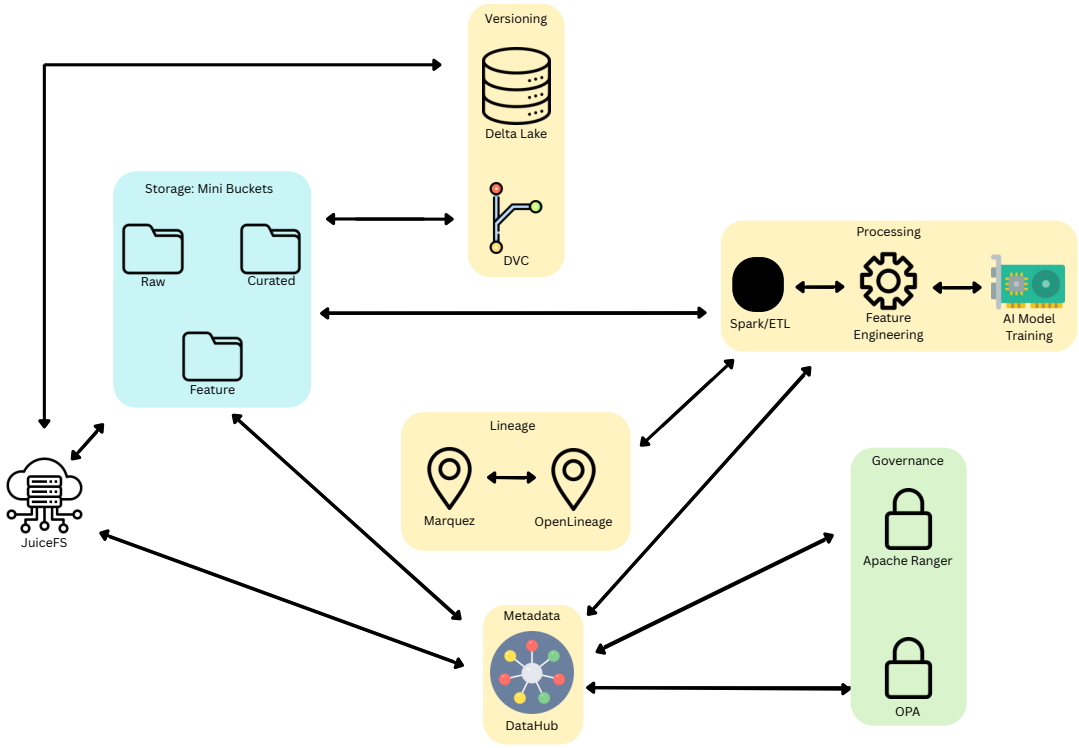


Figure 13 End-to-end data governance pipeline showing flow from raw ingestion through curated and feature layers. MinIO-backed JuiceFS storage provides the foundation, with Spark ETL pipelines processing data while emitting lineage events captured by OpenLineage and Marquez. DataHub maintains the metadata catalog, while Apache Ranger and OPA enforce governance policies at promotion boundaries. Delta Lake and DVC provide versioning and reproducibility guarantees throughout.

9.2 Metadata Cataloging and Taxonomic Organization

Metadata cataloging through DataHub transforms raw storage into a searchable, navigable knowledge graph where every asset is annotated with rich descriptive, operational, and governance-aligned metadata. DataHub serves as the central catalog organizing not merely individual documents but the full corpus topology, capturing relationships between source collections, processed datasets, derived features, and trained models in a unified graph structure. Every asset in our corpus is annotated with a comprehensive metadata schema spanning multiple dimensions simultaneously. Domain annotations classify content into categories including Agriculture, Culture, Education, News, Business, Healthcare, Sports, Law, Governance, Tourism, and BFSI (Banking, Financial Services, and Insurance), enabling construction of domain-specific subsets for targeted pretraining or evaluation. Language and script metadata distinguish between related but distinct linguistic varieties: Hindi in Devanagari script versus Hindi transliterated to Latin script versus Urdu

in Perso-Arabic script, while also capturing code-mixed content where multiple languages appear within single documents. Modality annotations identify whether content consists of plain text, PDF documents requiring OCR, images with embedded text, audio transcriptions, or code mixed with natural language documentation.

Quality tier metadata encodes multiple dimensions of data fidelity, including OCR confidence scores for digitized documents, readability metrics assessing linguistic complexity, and completeness indicators flagging truncated or corrupted content. License and sensitivity tags ensure safe promotion for downstream use by explicitly tracking intellectual property restrictions, personally identifiable information, and content requiring special handling due to cultural sensitivity or regulatory constraints. Source provenance captures origin information, which institutional archive, web domain, or digitization project contributed each document, enabling attribution and supporting partnerships with data providers who require usage tracking. Stage metadata indicates each asset's position in the processing pipeline, distinguishing raw ingested material from cleaned corpus ready for training from experimental features under development. Lineage metadata links assets to their processing history, capturing complete transformation graphs that enable reproducibility and debugging.

This rich metadata infrastructure serves multiple critical functions beyond basic organization. First, it enables precise corpus composition for targeted training objectives: constructing a legal domain corpus requires selecting by domain tag while filtering by quality tier and license compatibility. Second, it supports fairness and representation analysis by enabling quantitative assessment of corpus composition across languages, domains, and sources, revealing when certain linguistic communities or topical areas are underrepresented. Third, it facilitates automated dataset card generation, producing comprehensive documentation that satisfies emerging best practices for dataset transparency and responsible AI development. Fourth, it provides the substrate for governance policy enforcement, as policies can reference metadata predicates to conditionally permit or deny operations based on asset characteristics. DataHub's web interface transforms this metadata into powerful search and discovery capabilities, enabling researchers to navigate the corpus through faceted browsing, full-text search across metadata fields, and visual exploration of lineage graphs that reveal data provenance and downstream usage patterns.

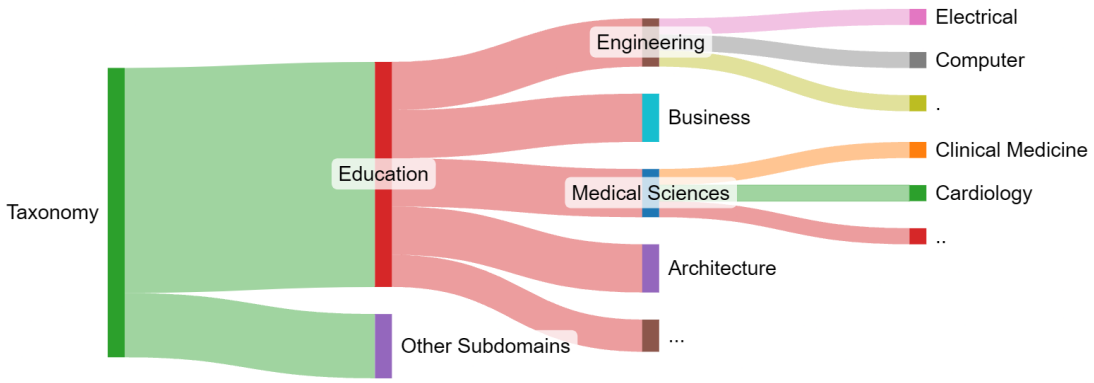


Figure 14 Sample Taxonomy Division of Education (1 of 12 Broad Domains)

Complementing the metadata infrastructure, our comprehensive taxonomy spanning over 1400 domains provides consistent structure and semantic coverage across the full range of tasks, languages, and knowledge areas represented in MILA. This taxonomy serves as a controlled vocabulary for domain classification, ensuring consistent annotation across different sources and processing stages while providing hierarchical organization that captures both broad categories and fine-grained specializations. A representative subset of the taxonomy appears in the appendix, illustrating how domain categories are organized into hierarchies—Healthcare branches into Ayurveda, Allopathic Medicine, Public Health, Traditional Healing Practices, and Medical Ethics, each with further subdivisions capturing specialized subdomains. This taxonomic structure enables not only consistent classification but also intelligent corpus subset construction: researchers can select content at varying levels of granularity, from all healthcare-related material to specifically Ayurvedic pharmaceutical texts, with the taxonomy automatically including appropriate subcategories. The taxonomy also guides synthetic data generation by ensuring comprehensive domain coverage in persona creation and providing structured prompts that elicit domain-appropriate reasoning patterns and knowledge.

9.3 Governance Policy Enforcement and Compliance

The governance layer operationalizes metadata and lineage into enforceable policies that mitigate legal, ethical, and quality risks while supporting controlled experimentation. We adopt a two-tier design: Apache Ranger manages role-based access, masking, and audit logging at scale, while Open Policy Agent (OPA) enforces fine-grained, context-aware rules. Ranger defines which teams can access which zones, the operations permitted on asset classes, and conditions for data promotion, for example, allowing only quality-approved, license-compliant assets to move from Curated to Feature Store. OPA implements these rules through programmable Rego policies, evaluating decisions at key points such as ingestion, transformation, feature sampling, and artifact publication. This policy-as-code approach enables versioning, automated testing, and auditable logs of all enforcement decisions.

License governance is particularly critical for Indic corpora, which span government releases, copyrighted works under agreements, web scrapes of uncertain status, and user-generated content with varied terms. Each source carries explicit license metadata, with conservative inheritance rules applied across transformations so that derived data inherits the most restrictive license. Policies prevent mixing of incompatible licenses and automatically filter datasets to license-compatible subsets. Compliance reports and audit trails document licensing status and guarantee that downstream models are trained only on permitted materials. Privacy protections address risks from large-scale crawls and user content. Multi-stage PII detection combines pattern matching, named entity recognition, and heuristic filters, triggering policies that range from full exclusion to redaction or metadata flagging for review. Domain-specific handling tailors enforcement: medical records require stricter filtering than news, social media identifiers demand careful redaction, and legal texts balance privacy with preservation of case law.

9.4 Versioning, Reproducibility, and Production Operations

The versioning and reproducibility layer ensures scientific rigor and production reliability for a continuously evolving corpus. Delta Lake provides ACID transactions and time travel, enabling atomic commits, rollbacks, and point-in-time queries that reconstruct historical corpus states. Each modification to curated or feature store data generates a new version capturing both the changes and metadata describing transformation logic, configuration, and execution context. This allows precise reconstruction of any prior corpus state for replication or analysis of how changes affect model behavior. Data Version Control (DVC) extends versioning to the full pipeline, tracking data artifacts alongside code, configuration, and dependencies. Integrated with Git, DVC enables collaborative workflows, branching experiments, and environment reconstruction without storing petabyte-scale data in Git itself. Researchers can retrieve specific pipeline versions, including processing code and corresponding data pointers, facilitating systematic experimentation with OCR heuristics, translation ensembles, and filtering strategies while preserving reproducibility.

Figure 13 illustrates the integrated workflow. Raw data enters via JuiceFS into the Raw zone with minimal processing. Apache Spark ETL pipelines apply cleaning, standardization, and quality filters while emitting OpenLineage events. DataHub builds lineage graphs enriched with processing provenance, and OPA enforces governance policies for promotion to the Curated zone based on quality, license, and domain balance. Curated data supports feature engineering, producing tokenized sequences, instruction tuning pairs, or domain-classified samples. Feature Store assets undergo final validation before training, with policies ensuring balanced sampling, exclusion of low-quality or sensitive material, and license compliance. Delta Lake captures atomic commits at every stage, DVC tracks code and configuration, and Marquez maintains the complete lineage from raw sources to final models. This architecture ensures that every token in MILA's 7.5 trillion token corpus has traceable provenance, verifiable quality, documented licensing, and reproducible processing history, transforming the data collection into a rigorously managed resource suitable for scientific investigation and production deployment. By embedding governance, quality, and reproducibility by design, the pipeline underpins both compliance and the scientific validity of downstream model evaluations.

10 Human-in-the-Loop Linguistic Validation

A critical insight emerging from our work on MILA is that achieving true linguistic quality for low-resource Indic languages requires moving beyond automated metrics and model-driven evaluation to systematically incorporate expert human judgment throughout the data curation pipeline. While automated evaluation provides essential scalability—enabling assessment of billions of tokens—it fundamentally cannot capture the subtle dimensions of linguistic naturalness, cultural appropriateness, and contextual fidelity that distinguish genuinely high-quality Indic language data from mechanically correct but culturally hollow text. This limitation proves particularly acute for low-resource

languages where automated metrics are themselves calibrated on inadequate reference corpora, potentially rewarding outputs that conform to limited training distributions while penalizing linguistically rich variations that fall outside narrow statistical norms. The challenge extends beyond surface-level grammaticality to encompass deeper questions of register appropriateness, dialectal variation, cultural resonance, and the preservation of linguistic features that automated systems—trained predominantly on high-resource languages—may incorrectly flag as errors.

Our approach integrates rigorous human-in-the-loop linguistic validation across all major pipeline components: OCR postprocessing, synthetic data generation, translation, and data distillation. Native language experts and trained linguists evaluate outputs iteratively across multiple complementary dimensions that together capture linguistic quality more comprehensively than any single metric. Fluency assessment examines whether text reads naturally and smoothly without awkward phrasing, unnatural sentence structures, or constructions that betray mechanical generation. Adequacy evaluation verifies that translated or generated content fully captures intended meanings without omissions, additions, or semantic distortions that alter propositional content. Grammar analysis identifies syntactic errors, morphological inconsistencies, and agreement violations that undermine linguistic correctness. Tone assessment ensures that register, formality level, and stylistic choices align appropriately with content type and target audience. Vocabulary richness evaluation measures whether texts employ diverse, expressive lexical choices rather than repetitive, simplified vocabulary characteristic of poor-quality synthetic data. Cultural appropriateness checking identifies content that, while linguistically correct, employs cultural references, examples, or framing inconsistent with Indian contexts. Readability assessment determines whether average speakers of each language can easily comprehend the text without specialized training or unusual linguistic sophistication.

This multidimensional evaluation framework enables identification of failure modes invisible to automated metrics. A translation might achieve high BLEU scores through literal word-for-word correspondence while producing text that native speakers judge unnatural or culturally inappropriate. Conversely, a high-quality translation employing idiomatic expressions and culturally appropriate adaptations might score lower on automated metrics precisely because it deviates from literal correspondence in service of naturalness. By systematically collecting expert judgments across these dimensions, we identify which processing pipelines, model configurations, and postprocessing strategies produce genuinely high-quality outputs for each language rather than merely optimizing automated metrics that may not align with human quality perceptions. Low-quality outputs flagged through this evaluation process are not discarded but rather corrected and reintegrated through iterative refinement loops, with pipelines rerun and configurations adjusted until consistently high scores across all dimensions are achieved for each language and task combination.

10.1 Quantitative Pipeline Selection Through Human-Calibrated Metrics

The practical value of human-centered evaluation becomes evident when comparing alternative processing pipelines to select optimal configurations for each language. Figure 15 presents a representative case study comparing two translation approaches: Mistral-24B-Instruct [24] versus an ensemble combining IndicTrans2 [27] and NLLB [67] across multiple Indic languages using readability as an illustrative metric. The results reveal dramatic performance heterogeneity across languages: Mistral-24B-Instruct excels for Assamese achieving 84 percent readability, Bengali at 70 percent, and Hindi at 76 percent, while the IndicTrans2-NLLB ensemble demonstrates superior performance for English at 84 percent, Gujarati reaching 98 percent, and Telugu at 88 percent. This language-specific performance variation reflects fundamental differences in training data availability, script complexity, and morphological richness across languages, validating our decision to employ pipeline selection strategies rather than applying uniform processing to all languages.

Such comparisons guide model selection for each language, ensuring outputs are not only syntactically correct but also culturally and contextually aligned with native speaker expectations. The evaluation process extends well beyond readability scores to encompass comprehensive assessment across all quality dimensions. For OCR outputs, evaluators assess not only character-level accuracy but whether reconstructed text preserves semantic coherence across line breaks, whether ligature decomposition produces linguistically valid sequences, and whether layout analysis correctly identifies reading order in multi-column documents with embedded figures. For translation, assessment examines whether target language outputs preserve subtle pragmatic meanings, maintain appropriate register and formality levels, and employ vocabulary natural to the domain rather than literal dictionary translations that sound stilted. For synthetic data generation, evaluation verifies that persona-driven outputs genuinely reflect Indian cultural contexts rather than superficially adapted Western content, that reasoning patterns align with domain-specific norms within Indian professional and academic contexts, and that generated examples employ culturally appropriate scenarios and references.

The scoring process operates through iterative refinement loops where low-quality outputs are not merely flagged but

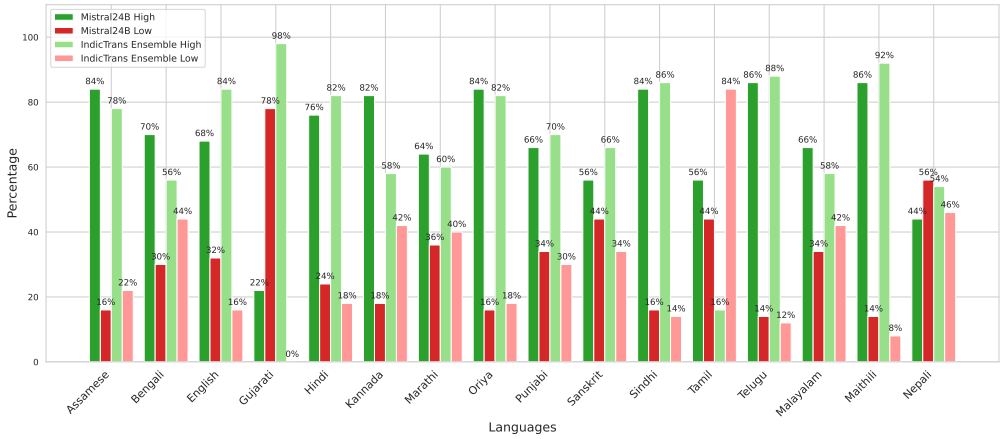


Figure 15 Readability comparison between Mistral-24B-Instruct and IndicTrans2-NLLB ensemble across Indic languages, demonstrating language-specific performance heterogeneity that motivates adaptive pipeline selection. Each language achieves optimal results with different model configurations, with Mistral excelling for Assamese, Bengali, and Hindi while the ensemble performs better for English, Gujarati, and Telugu.

actively corrected by linguists, with corrections fed back to improve processing pipelines. When evaluation reveals systematic errors, such as consistent mistranslation of domain-specific terminology, inappropriate register choices for particular content types, or recurring grammatical patterns characteristic of mechanical translation, these patterns inform targeted improvements to models, prompts, and postprocessing rules. Multiple evaluation rounds continue until all metrics consistently meet high standards, with each iteration incorporating lessons from previous rounds to progressively elevate quality. This ensures that each language and task combination ultimately leverages a specialized, validated pipeline preserving linguistic integrity, cultural context, and factual fidelity rather than accepting mediocre outputs from generic processing approaches. The result is a corpus where quality is not assumed based on automated metrics but actively verified and refined through expert judgment, with continuous improvement driven by systematic analysis of failure modes and targeted interventions addressing identified weaknesses.

10.2 Structured Evaluation Protocols and Criteria Standardization

The effectiveness of human evaluation depends critically on providing evaluators with clear, standardized criteria and systematic protocols that ensure consistency across annotators, languages, and evaluation rounds. We developed comprehensive evaluation guidelines that operationalize abstract quality dimensions into concrete assessment procedures with explicit decision rules and illustrative examples. Evaluators receive detailed instructions specifying how to identify and categorize different error types, what severity levels to assign based on impact on comprehensibility and naturalness, and how to provide actionable feedback enabling targeted corrections. This standardization proves essential for maintaining evaluation reliability as the project scales across 16 languages with different evaluator teams, ensuring that a score of 4 out of 5 for fluency carries consistent meaning whether applied to Assamese translations or Telugu synthetic data.

The evaluation protocol structures assessment around clearly defined, answerable questions for each quality dimension. As illustrated in Table 22, for instance, fluency and readability are evaluated by asking whether the translation reads naturally and smoothly in the target language, without awkward phrasing, unnatural sentence structures, or grammatical errors. Ratings are then assigned on a scale from 1 (very unnatural) to 5 (perfectly natural). Similar rating procedures are applied across all other evaluation criteria, and the resulting scores are subsequently used to provide feedback for enhancing the translation pipeline.

Beyond numerical ratings, evaluators provide critical written feedback detailing specific issues identified in assessed samples. This qualitative feedback proves invaluable for diagnosing systematic problems and guiding targeted improvements. Representative feedback examples from our evaluation process illustrate the types of insights human judgment provides that automated metrics miss entirely. One evaluator noted that "Hindi translation is not proper, uses overly formal Sanskritized vocabulary inappropriate for the conversational tone of the source text," identifying a register mismatch invisible to most automated metrics. Another flagged content as "not depicting true picture, contains anti-

Evaluation Criterion	Description
Fluency & Readability	Does the translation read naturally and smoothly in your language, without awkward phrasing, unnatural sentence structures, or grammatical errors?
Adequacy & Meaning Preservation	Does the translated text fully capture the meaning of the original English sentence without omitting, adding, or distorting information?
Use of Rich & Appropriate Vocabulary	Does the translation use a rich and diverse vocabulary that feels natural and expressive in your language?
Cultural & Contextual Appropriateness	Are there any cultural inconsistencies, unnatural phrases, or word choices that feel out of place or confusing in your language?
Grammar & Sentence Structure	Are the grammar, syntax, and sentence structures well-formed and correct in your language?
Consistency & Tone Matching	Does the translation maintain the same tone, formality, and style as the original English text?
Readability & Ease of Understanding	Is the translation easy to read and understand for an average speaker of your language?

Table 22 Sample Translation Quality Evaluation Criteria with Descriptions

national sentiment," catching politically sensitive framing that requires cultural knowledge to identify. Concerns about regional bias appeared in feedback like "North-South divide should be avoided" and "Why mention a particular state? Shows regional bias," ensuring generated examples don't inadvertently reinforce stereotypes. Terminology choices received scrutiny: "Instead of 'bhedbhaav' it should be 'indifference', passage tone should suggest solutions rather than expressing anger, needs softer and more polite terminology." Factual accuracy checking emerged in feedback such as "Doesn't seem to be reality—fact check percentage cited" and "Are these statistics verified?" Authenticity concerns surfaced in notes like "Indian name pronunciation issues—proper authentic usage of Indian vocabulary should be present."

This feedback directly informs pipeline improvements through systematic categorization and analysis. Common feedback patterns indicate where models consistently struggle—such as inappropriate register choices, regional bias, or factual inaccuracies—enabling targeted interventions. For translation pipelines, feedback revealing consistent terminology issues for specific domains motivates development of domain-specific glossaries and constraints. For synthetic generation, feedback identifying cultural inappropriateness guides refinement of persona specifications and generation prompts to better capture Indian contexts. The evaluation framework implements a zero-data-loss policy where low-quality data is corrected and updated based on feedback rather than simply discarded, ensuring continuous improvement while maximizing the value extracted from expensive human annotation. Multiple evaluation rounds with iterative refinement continue until outputs consistently achieve high scores across all dimensions, with each iteration incorporating lessons from previous feedback to progressively eliminate failure modes.

10.3 Addressing Dialectal Variation and Practical Usability

Standard evaluation approaches fall short when targeting non-urban, monolingual populations who speak dialects diverging from formal standard varieties. Evaluations conducted by native speakers of standard dialects—typically taught in schools and represented in digital corpora, do not reflect whether dialect speakers of varying literacy levels can understand or use model outputs. This limitation is especially consequential for agricultural advisory systems, where users predominantly speak regional dialects with vocabulary, pronunciation, and grammatical features absent from standard Hindi. A system generating flawless standard Hindi may be incomprehensible or culturally alien to a Bhojpuri-speaking farmer, whereas outputs incorporating dialectal features might score lower on standard metrics precisely because they deviate from formal norms. Indic languages exhibit profound dialectal diversity that conventional corpora and evaluation methods erase. Hindi alone varies dramatically across regions: Haryanvi, Punjabi-influenced Hindi, Bihari, and Jharkhandi dialects differ in vocabulary, morphology, and syntax, particularly for everyday agricultural objects and practices. Folk terms for vegetables, clothing, tools, and farming processes carry rich pragmatic and cultural associations, learned orally rather than through formal education. Standard language models trained primarily on written corpora lack this dialectal vocabulary and cultural grounding. Even when dialectal text is included, the dominance of

standard forms causes models to favor these over less frequent dialectal alternatives, creating a disconnect between model outputs and the needs of rural users.

For example, most farmers in target demographics are illiterate or semi-literate, with even literate individuals preferring folk vocabulary over standard language in practical agricultural discussions. Furthermore, farmers express strong preference for speech-based interaction over text, reflecting both literacy constraints and the practical reality that speech is more natural and efficient for real-time agricultural decision-making. Text-based systems, regardless of linguistic quality, exclude large portions of the target population, while speech interfaces employing dialectal vocabulary and natural prosody enable genuine accessibility. To address this gap, we propose targeted evaluation assessing dialectal appropriateness and practical usability for non-urban populations, starting with pilot experiments before broader deployment. The pilot focuses on agriculture and two Hindi dialects, including Bhojpuri, chosen for its wide geographic spread, rich folk vocabulary, and large farming population. Native Bhojpuri-speaking agricultural workers serve as evaluators, judging whether model outputs use vocabulary, grammar, and cultural references natural to their variety rather than imposing formal Hindi.

Evaluation extends beyond standard quality dimensions to capture real-world usability. Vocabulary naturalness considers folk terms for crops, tools, and processes. Cultural resonance ensures examples and scenarios reflect rural lived experience. Comprehensibility for low-literacy users evaluates sentence structures and discourse organization suitable for limited formal education. Speech interface suitability examines whether outputs would sound natural when rendered orally with local pronunciation and prosody. These criteria complement standard measures of grammar and factual accuracy. Beyond agriculture, dialectal evaluation redefines linguistic quality for low-resource language technology. Conventional frameworks privilege formal, written varieties, marginalizing dialects spoken by millions. Incorporating dialectal variation fosters inclusive systems that respect linguistic diversity and folk registers systematically erased by standard corpora and automated metrics. Human-in-the-loop evaluation thus becomes essential for ensuring MILA and related systems reflect India's full linguistic richness.

Crucially, building a high-quality Indic multilingual dataset relied on rigorous human-in-the-loop validation across OCR, synthetic generation, translation, and data distillation. Native experts iteratively evaluated fluency, adequacy, grammar, tone, vocabulary richness, cultural appropriateness, and readability. Low-quality outputs were corrected and reintegrated, with pipelines rerun until consistently high scores were achieved, ensuring optimal quality for every language and task.

11 Final Experiment

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 ([15]) pretrained checkpoint and measure its Indic MMLU ([22]) 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) [76], BLOOMZ-7B [39], LLaMA-2-7B [71], Gemma-7B ([64, 65, 66]), Mixtral-7B [25], and Granite-7B [37]. Results for these models are provided in the appendix. Together, these evaluations demonstrate both the effectiveness and fairness gains enabled by MILA.

Architecture attributes	Values
Model Architecture	Qwen3-600M (causal-language-model)
Hidden size	1536
Intermediate size (FFN)	6144
Max Position Embeddings	2048
Num of Attention Heads	16
Num of Hidden Layers	24
Num of Query Groups	16
Normalization	RMSNorm
Activation Function	swiglu
Attention Type	Multi-head Attention with RoPE
Position Embedding Type	RoPE (rotary)
Dropout (hidden/attn/ffn)	0.0 / 0.0 / 0.0
Precision	bf16 (mixed)

Table 23 Architecture Details of QWEN3-600M

11.1 Results

Table 24 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.

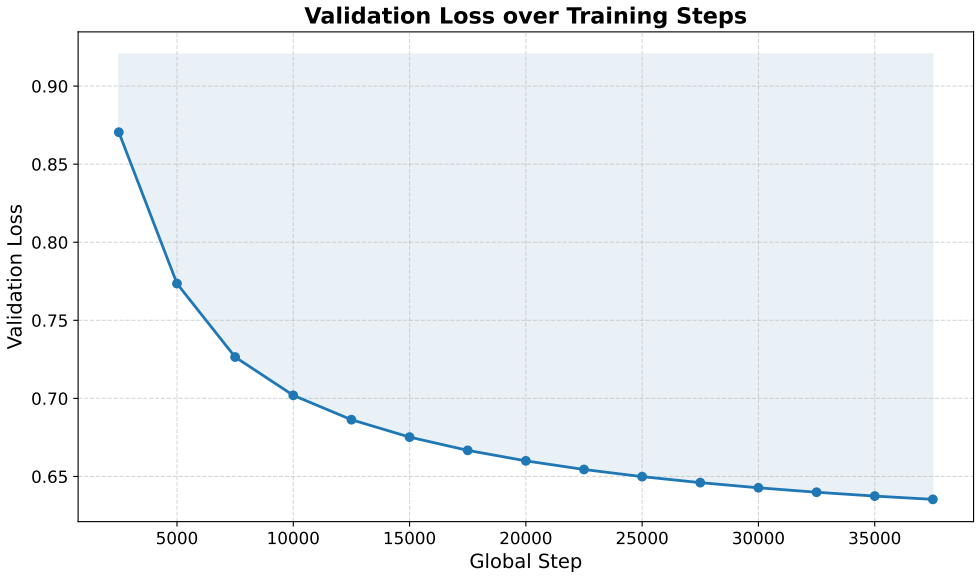


Figure 16 Validation Loss Plot for Qwen3-600M on MILA

Table 24 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 25). 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

Table 25 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

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.

12 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

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

[2] 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. https://arxiv.org/abs/2309.16609.

[3] 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, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. https://aclanthology.org/2020.acl-main.417/.

[4] 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. https://arxiv.org/abs/2005.14165.

[5] 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.

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[7] Common Crawl Foundation. Common crawl dataset. https://commoncrawl.org/.

[8] 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. http://arxiv.org/abs/1911.02116.

[9] 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. https://arxiv.org/abs/2508.10975.

[10] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124, 2019.

[11] 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. https://arxiv.org/abs/2412.19437.

[12] 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. http://arxiv.org/abs/1810.04805.

[13] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. https://aclanthology.org/2023.acl-long.693/.

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

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

[16] Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, 2023. https://arxiv.org/abs/2305.16307.

[17] 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. https://arxiv.org/abs/2101.00027.

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

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

[20] 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. https://doi.org/10.1162/tacl_a_00474.

[21] 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. https://arxiv.org/abs/2407.21783.

[22] 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. https://arxiv.org/abs/2009.03300.

[23] 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, page 2206–2222. ACM, June 2022. doi: 10.1145/3531146.3534637. http://dx.doi.org/10.1145/3531146.3534637.

[24] 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. https://arxiv.org/abs/2310.06825.

[25] 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. https://arxiv.org/abs/2401.04088.

[26] 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. https://arxiv.org/abs/2411.04257.

[27] 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)*, page 15831–15879. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.843. http://dx.doi.org/10.18653/v1/2024.acl-long.843.

[28] 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. https://arxiv.org/abs/2303.03915.

[29] 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, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. https://aclanthology.org/2022.acl-long.577/.

[30] 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. https://arxiv.org/abs/2406.11794.

[31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[32] 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.

[33] 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.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

[36] 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. https://arxiv.org/abs/2505.02009.

[37] 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. https://arxiv.org/abs/2405.04324.

[38] Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023.

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

[40] Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, et al. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*, 2025.

[41] 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. https://arxiv.org/abs/2309.09400.

[42] 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. https://arxiv.org/abs/2406.11704.

[43] 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. https://arxiv.org/abs/2508.10925.

[44] 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, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

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

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

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

[48] 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.

[49] 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. https://arxiv.org/abs/2506.20920.

[50] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

[51] Kundeshwar Pundalik, Piyush Sawarkar, Nihar Sahoo, Abhishek Shinde, Prateek Chanda, Vedant Goswami, Ajay Nagpal, Atul Singh, Viraj Thakur, Vijay Dewane, et al. Param-1 bharatgen 2.9 b model. *arXiv preprint arXiv:2507.13390*, 2025.

[52] 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. https://arxiv.org/abs/2412.15115.

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

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

[55] 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. https://arxiv.org/abs/2112.11446.

[56] 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. https://arxiv.org/abs/1910.10683.

[57] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10: 145–162, 2022. doi: 10.1162/tacl_a_00452. https://aclanthology.org/2022.tacl-1.9/.

[58] 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. http://arxiv.org/abs/1911.04944.

[59] 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. https://arxiv.org/abs/2405.01582.

[60] 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. https://arxiv.org/abs/2309.10818.

[61] 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. https://arxiv.org/abs/2404.16816.

[62] 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. https://arxiv.org/abs/2402.00159.

[63] 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.

[64] 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, 2024. https://arxiv.org/abs/2403.08295.

[65] 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, 2024. https://arxiv.org/abs/2408.00118.

[66] 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. https://arxiv.org/abs/2503.19786.

[67] 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 behind: Scaling human-centered machine translation, 2022. https://arxiv.org/abs/2207.04672.

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

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

[70] 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, 2023. https://arxiv.org/abs/2302.13971.

[71] 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, 2023. https://arxiv.org/abs/2307.09288.

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

[73] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan

He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[74] 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. https://arxiv.org/abs/2411.12372.

[75] Wikimedia Foundation. Wikipedia dumps. https://dumps.wikimedia.org/.

[76] 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. https://arxiv.org/abs/2010.11934.

[77] 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. https://arxiv.org/abs/2407.10671.

[78] 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*, pages 656–668, 2024.

[79] 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. https://arxiv.org/abs/2211.15649.

[80] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[81] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

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

[83] Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. Hunyuan-mt technical report, 2025. https://arxiv.org/abs/2509.05209.

A MILA Evaluation Prompt: Math

Evaluation Prompt: Math

Context

You are a Math Expert with specialization in Abstract Algebra, College Mathematics, Elementary Mathematics, Formal Logic, Econometrics, High School Mathematics, and High School Statistics, holding a BS in Chemical Engineering from IIT. With deep expertise in mathematical reasoning and quantitative logic, they are adept at identifying whether a translated text preserves technical accuracy, mathematical terminology, symbols, and conceptual precision. They carefully check if the mathematical intent of the original English text is faithfully represented in the translated version. Their evaluation style is structured, analytical, and strictly objective, assigning a rating from 1 to 10 for mathematical correctness and fidelity.

Objective

Evaluate the Translated Question compared to the Original Question in English. Focus only on mathematical accuracy: Are all numbers, formulas, symbols, and equations correctly preserved? Is the logical and mathematical structure consistent with the original?

Inputs:

Original Question (English):
`"{0}"`
`"{1}"`

Translated Question:
Language of Translation: `"{3}"`
`"{2}"`

Output Format (strict JSON):

Give a single integer rating from 1 to 10 (1 = worst, 10 = excellent).

```
{
  "Math Expert": <rating>
}
```

B Translation Benchmark Results

B.1 Evaluation of Baseline MT and LLMs on Indic Languages

We evaluated the general translation capabilities of current open-source models on Indic languages using **ai4bharat/IN22-Gen** and **google/IndicGenBench_flores_in**. Both baseline MT systems and large language models (LLMs) were tested using CHRF, CHRF++, and SACREBLEU metrics.

Results show that LLMs generally provide more fluent and context-aware translations, especially for morphologically rich languages, while baseline MT models perform well for high-resource languages but lag on low-resource or complex languages. Performance varies across language pairs, highlighting the uneven support for Indic languages in current open-source models.

B.1.1 Results for ai4bharat/IN22-Gen

B.1.2 Results for google/IndicGenBench$_{flores_in}$

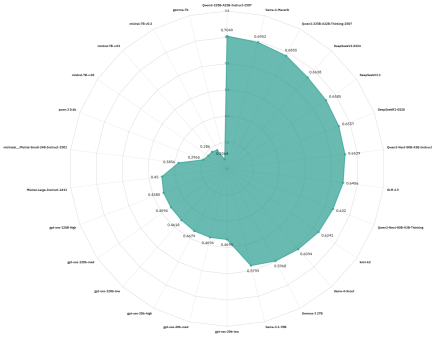
Figure 17 Evaluation results for ai4bharat/IN22-Gen using open-source models, assessed with chrF, chrF++, and sacreBLEU metrics – part I

Figure 18 Evaluation results for ai4bharat/IN22-Gen using open-source models, assessed with chrF, chrF++, and sacreBLEU metrics – part II

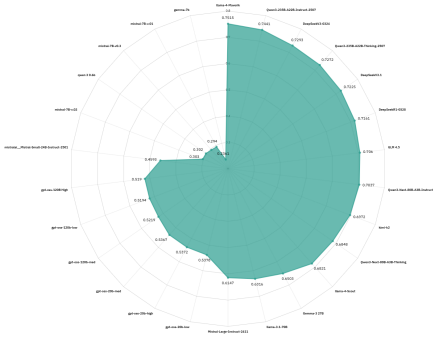
Figure 19 Evaluation results for ai4bharat/IN22-Gen using open-source models, assessed with chrF, chrF++, and sacreBLEU metrics – part III

Figure 20 Evaluation results for ai4bharat/IN22-Gen using open-source models, assessed with chrF, chrF++, and sacreBLEU metrics – part IV

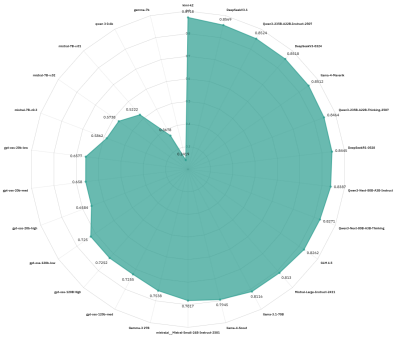
Figure 21 Evaluation results for ai4bharat/IN22-Gen using open-source models, assessed with chrF, chrF++, and sacreBLEU metrics – part V



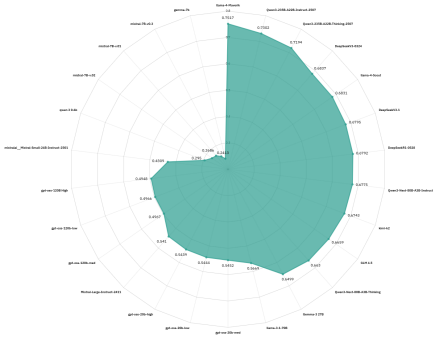
(a) Assamese



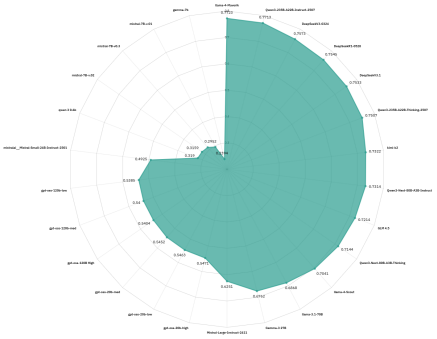
(b) Bengali



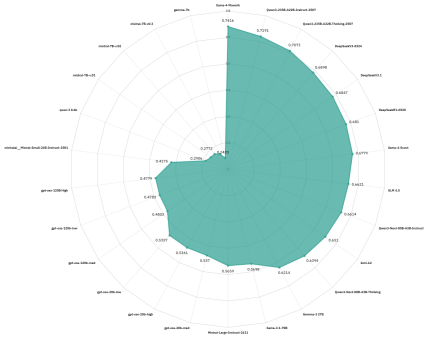
(c) English



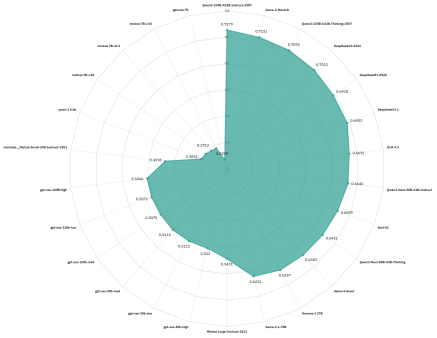
(d) Gujarati



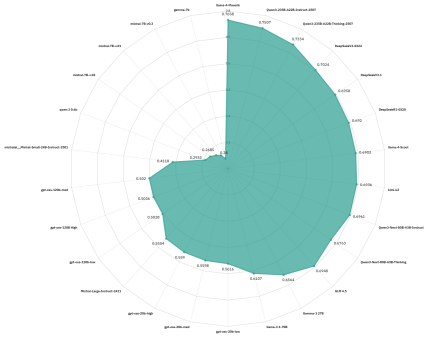
(e) Hindi



(f) Kannada

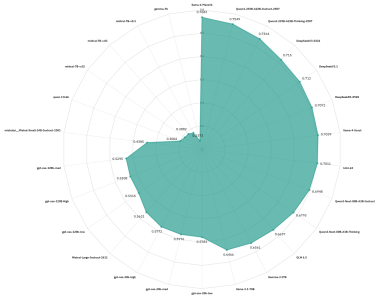


(g) Maithili

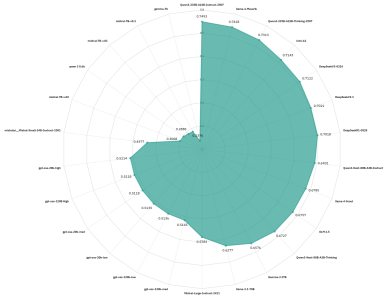


(h) Malayalam

Figure 22 Eight images arranged in two columns (4 rows).

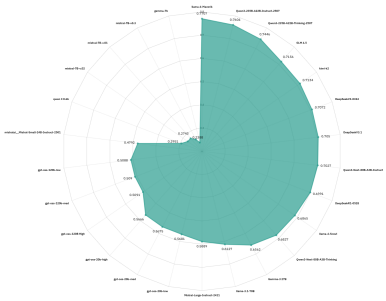


(a) Marathi

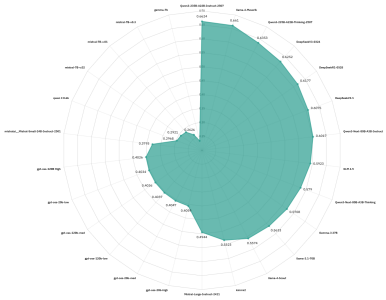


(b) Nepali

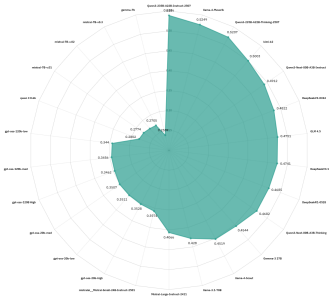
(c) Oriya



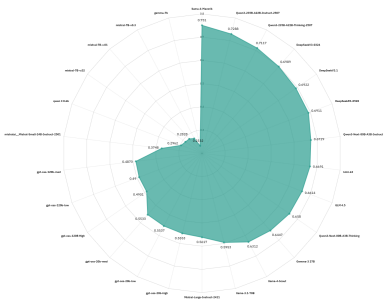
(d) Punjabi



(e) Sanskrit



(f) Sindhi

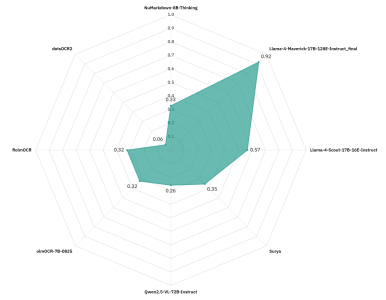


(g) Tamil

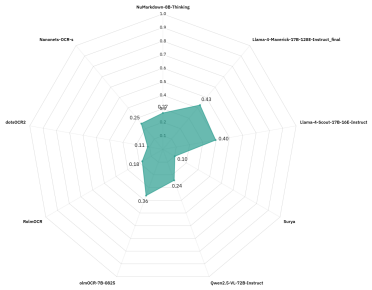
(h) Telugu

Figure 23 Eight images arranged in two columns (4 rows).

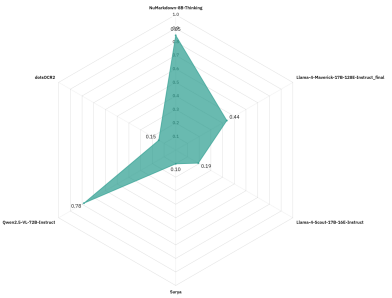
C OCR Benchmark Results



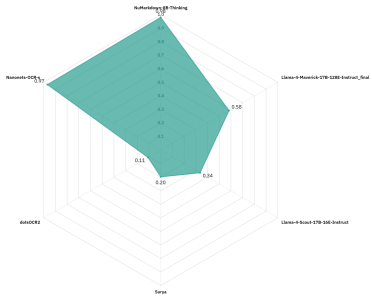
(a) Sanskrit



(b) Sindhi

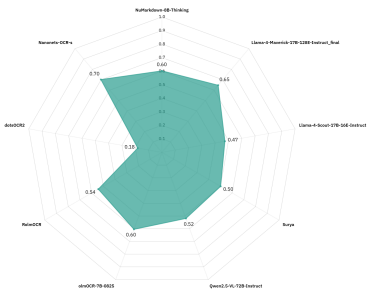


(c) Tamil

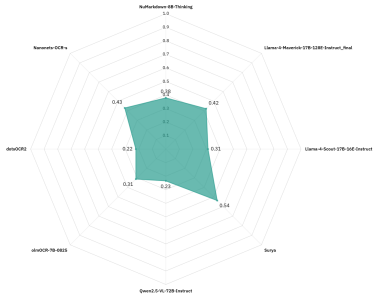


(d) Telugu

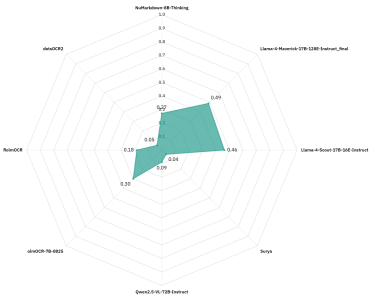
Figure 24 Eight images arranged in two columns (4 rows).



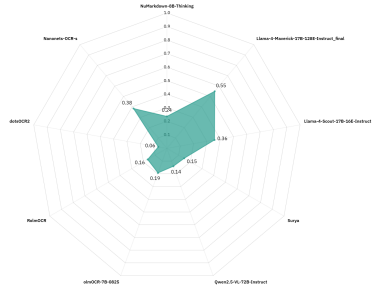
(a) Assamese



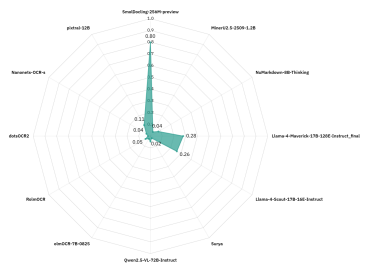
(b) Bengali



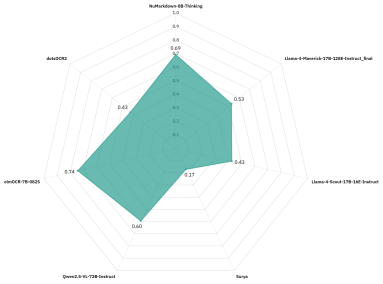
(c) Bodo



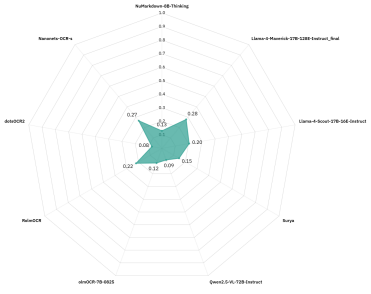
(d) Dogri



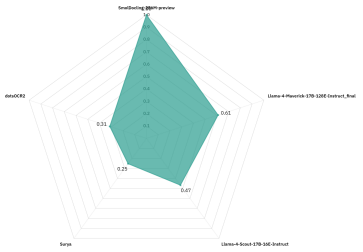
(e) English



(f) Gujarati

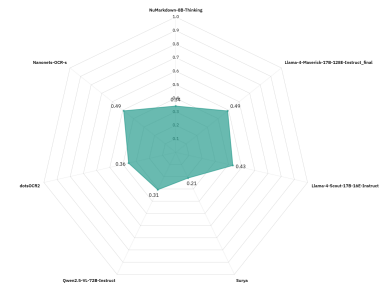


(g) Hindi

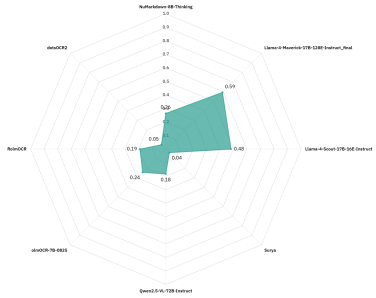


(h) Kannada

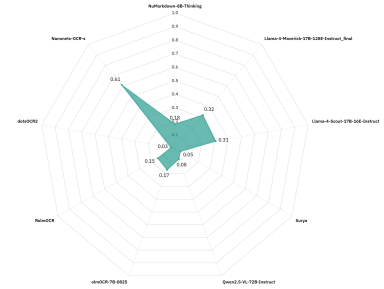
Figure 25 Eight images arranged in two columns (4 rows).



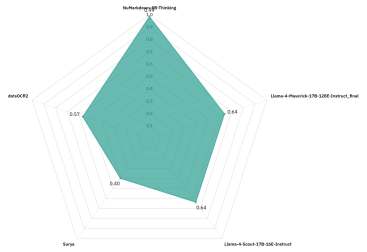
(a) Kashmiri



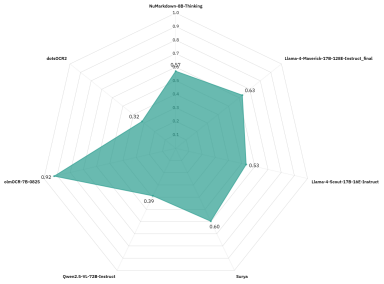
(b) Konkani



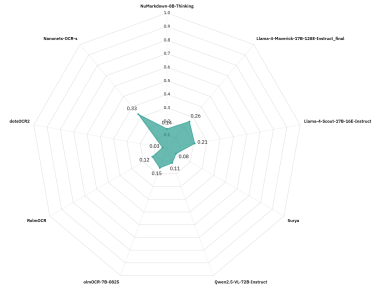
(c) Maithili



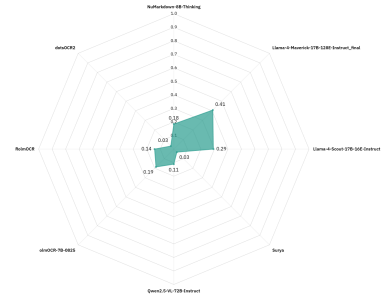
(d) Malayalam



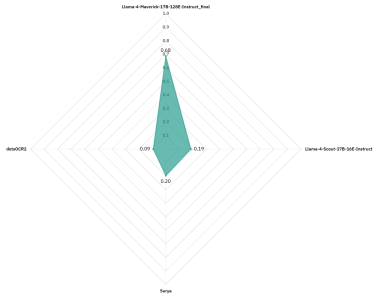
(e) Manipuri



(f) Marathi



(g) Nepali



(h) Punjabi

Figure 26 Eight images arranged in two columns (4 rows).