INDICLLMSUITE: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages

Mohammed Safi Ur Rahman Khan[1*] **Priyam Mehta**[1*] **Ananth Sankar**[1]
Umashankar Kumaravelan[1] **Sumanth Doddapaneni**[1,2] **Suriyaprasaad B**[1,4†]
Varun Balan G[1,5†] **Sparsh Jain**[1,6†] **Anoop Kunchukuttan**[1,2,3]
Pratyush Kumar[1,2,7] **Raj Dabre**[2,8] **Mitesh M. Khapra**[1,2‡]
[1] Nilekani Centre at AI4Bharat [2]Indian Institute of Technology, Madras [3]Microsoft
[4] SLIET [5]IIIT D&M Kancheepuram [6]MAIT [7]Sarvam AI [8]NICT, Japan

Abstract

Despite the considerable advancements in English LLMs, the progress in building comparable models for other languages has been hindered due to the scarcity of tailored resources. Our work aims to bridge this divide by introducing an expansive suite of resources specifically designed for the development of Indic LLMs, covering 22 languages, containing a total of 251B tokens and 74.8M instruction-response pairs. Recognizing the importance of both data quality and quantity, our approach combines highly curated manually verified data, unverified yet valuable data, and synthetic data. We build a clean, open-source pipeline for curating pre-training data from diverse sources, including websites, PDFs, and videos, incorporating best practices for crawling, cleaning, flagging, and deduplication. For instruction-fine tuning, we amalgamate existing Indic datasets, translate/transliterate English datasets into Indian languages, and utilize LLaMa2 and Mixtral models to create conversations grounded in articles from Indian Wikipedia and Wikihow. Additionally, we address toxicity alignment by generating toxic prompts for multiple scenarios and then generate non-toxic responses by feeding these toxic prompts to an aligned LLaMa2 model. We hope that the datasets, tools, and resources released as a part of this work will not only propel the research and development of Indic LLMs but also establish an open-source blueprint for extending such efforts to other languages. The data and other artifacts created as part of this work are released with permissive licenses at https://github.com/AI4Bharat/IndicLLMSuite

1 Introduction

Building Large Language Models (LLMs) is an inherently data-intensive process requiring a comprehensive set of resources for pre-training (Raffel et al., 2020; Xue et al., 2021; Gao et al., 2021; Penedo et al., 2023; Nguyen et al., 2023a; Abadji et al., 2022) and fine-tuning (Longpre et al., 2023; Conover et al., 2023; Köpf et al., 2023; Ding et al., 2023a). The last year has seen remarkable progress in building English LLMs, thanks to open-source models (Touvron et al., 2023a,b; Jiang et al., 2023, 2024a; Almazrouei et al., 2023) developed using comprehensive datasets containing such resources. Nonetheless, this progress has largely bypassed low and mid-resource languages due to the the lack of data resulting from the lack of open source pipelines for curating data for such languages from diverse sources such as websites (which require crawling and extraction), books (which require OCR) and videos (which require transcription). Further, for instruction fine-tuning, English LLMs now rely on model-generated data such as ShareGPT[1], Self-Instruct (Wang et al., 2023a), Evol-Instruct (Xu et al., 2023a), UltraChat (Ding et al., 2023a), etc. However, for low and mid resource languages this option is not available due to lack of high quality LLMs, leading to a chicken and egg problem, further widening the gap between the *haves* and the *have-nots*.

A case in point is that of languages from the Indian sub-continent which collectively are spoken by over 1.4 billion people. We focus on the 22 languages recognised in the 8th schedule of the Indian constitution. These languages, despite their significant number of speakers, receive minimal representation in the training datasets and tokenizers of current open-source LLMs (Touvron et al., 2023b; Jiang et al., 2024a; Almazrouei et al., 2023) leading to a notable exclusion of their rich cultural contexts and nuances. In this work, we address this disparity by making the following contributions:

1. SANGRAHA: Pretraining data containing 251B tokens[2] summed up over 22 languages extracted

* Equal Contribution.
† Work done during an internship at AI4Bharat
‡ Corresponding Author: miteshk@cse.iitm.ac.in

[1] https://sharegpt.com/

[2] We built a custom tokenizer which supports English and

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15831–15879
August 11-16, 2024 ©2024 Association for Computational Linguistics

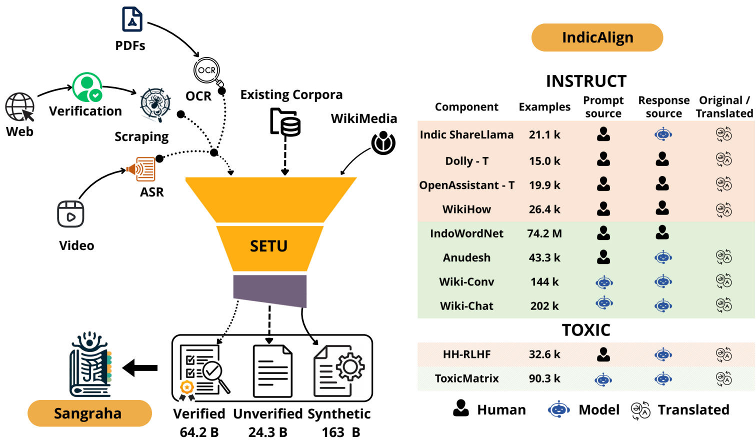


Figure 1: Overview of the different components present in INDICLLMSUITE.

from curated URLs, existing multilingual corpora, and large-scale translations.

2. SETU: Spark-based (Zaharia et al., 2016) distributed pipeline customised for Indian languages for extracting content from websites, PDFs and videos, with in-built stages for cleaning, filtering, toxicity removal and deduplication.

3. INDICALIGN - INSTRUCT: A diverse collection of 74.7 million prompt-response pairs across 22 languages collected through four methods: aggregating existing Instruction Fine-Tuning (IFT) datasets, translating English datasets into 14 Indian languages using an open-source translation model, creating context-grounded conversations from India-centric Wikipedia articles using open source LLMs, and establishing a crowdsourcing platform called *Anudesh* for prompt collection.

4. INDICALIGN - TOXIC: 123K pairs of toxic prompt and non-toxic responses generated using open source English LLMs and translated to 14 Indian languages for safety alignment of Indic LLMs.

We collectively refer to the above as INDICLLMSUITE. We try to balance quality and quantity while acknowledging recent trends of using synthetic data for building powerful LLMs for English (Gunasekar et al., 2023; Li et al., 2023c) as well as low resource languages (Nguyen et al., 2023b; Li et al., 2023b). To ensure quality, we take help from humans to verify websites to flag noisy or machine translated content and to create toxicity lists for Indian languages. On the other hand, to ensure explicit representation of prompt-response pairs grounded in Indian context we take the help of powerful open source LLMs to generate grounded conversations from India-centric Wikipedia articles. We recognize the need to represent diverse knowledge and alignment information in Indic languages for better performance of LLMs in Indic languages. Hence, we undertake large-scale machine translation of rich English resources like Wikipedia as well as English finetuning datasets into Indian languages using SOTA open-source MT models. We thus balance source original data with translated and LLM-generated data to create the above collection.

We believe that these choices can be replicated across other languages to create LLMSuites. All the code, tools and datasets developed as a part of this work will be publicly released and hopefully advance the development of LLMs for Indian languages. Given that LLM training is an expensive exercise, we plan to undertake community-effort to train LLMs, where multiple groups can pool together computing resources to build a high-quality Indic language LLM.

2 Related Works

We organise the Related Work into 3 sections in line with our main contributions.

Multilingual Datasets. Previous works like OSCAR (Abadji et al., 2022), CC100 (Conneau et al., 2020), and mC4 (Raffel et al., 2020) are curated from CommonCrawl dumps through extensive cleaning stages. MADLAD-400 (Kudugunta et al., 2023) extends to 419 languages, incorporating human audits and iterative refinement, along with language family-specific filters. ROOTS (Laurençon et al., 2023) and CulturaX (Nguyen et al., 2023a) combined existing datasets and used strict clean-

Indian languages and has an average fertility of 1.3 to 2.79 across the 22 languages. We use this tokenizer for all the reported statistics unless mentioned otherwise.

ing procedure to create data in 59 and 167 languages respectively. In contrast, we curate a list of verified websites for Indian languages and use a pipeline specifically built for Indian languages to clean up data from some of the existing collections listed above. In addition to HTML documents from web crawls, we also include PDFs and speech transcripts from a variety of sources.

Data Curation Pipelines. For scraping websites we rely on popular tools like Trafilatura (Barbaresi, 2021a) and jusText (Endrédy and Novák, 2013) which are widely used in recent works (Penedo et al., 2023; Gao et al., 2021). For language identification, in addition to popular tools such as, cld3[3], langdetect[4], fasttext (Wenzek et al., 2020), SSLID (Kudugunta et al., 2023), we use a custom Indic language LID tool (Madhani et al., 2023a). For ensuring high quality of the processed data recent works have proposed rule-based approaches (Raffel et al., 2020; Xue et al., 2021; Laurençon et al., 2023; Rae et al., 2021), approaches which rely on KenLM perplexity (Wenzek et al., 2020; Laurençon et al., 2023; Nguyen et al., 2023a) and machine learning-based approaches (Brown et al., 2020). In this work, we largely rely on rule based approaches. Similarly, for filtering toxic content we use word lists (Raffel et al., 2020), blocklists of URLs (Penedo et al., 2023), and language specific heuristics (Kudugunta et al., 2023). Lastly, deduplication (Broder, 1997; Charikar, 2002; Abbas et al., 2023) is important to reduce memorisation (Carlini et al., 2023) and improve LLM performance (Lee et al., 2022), especially at scale (Hernandez et al., 2022). In this work, we perform deduplication based on URLs, and fuzzy techniques like MinHash (Broder, 1997).

Supervised Fine-Tuning Datasets. Existing IFT datasets either contain human-human interactions such as Dolly (Conover et al., 2023) and OpenAssistant (Köpf et al., 2023), human-GPT interactions such as ShareGPT[5] and WildChat (Zhao et al., 2024) or LLM-generated instructions (Wang et al., 2023b) typically bootstrapped with a few prompts, such as Alpaca (Taori et al., 2023), WizardLM (Xu et al., 2023a), Camel (Li et al., 2023a), Ultrachat (Ding et al., 2023a), Baize (Xu et al., 2023c) amongst others. Multilingual instruction tuning datasets are created from English IFT datasets using machine translation of prompts and/or outputs (Li et al., 2023b; Wei et al., 2023). In this work, we translate existing IFT datasets to Indian languages and also create LLM-augmented conversations grounded in Indian context.

[3]https://github.com/google/cld3/
[4]https://github.com/shuyo/language-detection
[5]https://sharegpt.com/

Code	SV	SS	SU	Total Tokens
asm	292.14	11696.41	17.52	12006.0
ben	10604.46	13814.14	5608.89	30027.5
brx	1.5	-	-	1.5
doi	0.064	-	-	0.064
eng	12759.9	-	-	12759.9
gom	10.1	-	-	10.1
guj	3647.9	12934.5	597.0	17179.4
hin	12617.3	9578.7	12348.3	34544.3
kan	1778.3	12087.4	388.8	14254.5
kas	0.5	-	-	0.5
mai	14.6	-	-	14.6
mal	2730.8	13130.0	547.8	16408.6
mar	2827.0	10816.7	652.1	14295.8
mni	7.4	-	-	7.4
npi	1822.5	10588.7	485.5	12896.7
ory	1177.1	11338.0	23.7	12538.8
pan	1075.3	9969.6	136.9	11181.8
san	1329.0	13553.5	9.8	14892.3
sat	0.3	-	-	0.3
snd	258.2	-	-	258.2
tam	3985.1	11859.3	1515.9	17360.3
urd	3658.1	9415.8	1328.2	14402.1
tel	3706.8	11924.5	647.4	16278.7
Total	64306.1	162707.9	24307.7	**251321.0**

Table 1: Number of tokens (in Millions) in each split of Sangraha. (SV: SANGRAHA VERIFIED, SS: SANGRAHA SYNTHETIC, SU: SANGRAHA UNVERIFIED)

3 SANGRAHA

In this section we describe the composition and curation of SANGRAHA spanning verified (64B), unverified (24B) and synthetic (162B) content for a total of 251B tokens.

3.1 Data Composition

3.1.1 Sangraha Verified

To ensure high quality, we introduce the Sangraha Verified dataset, comprising Web Data, PDF Data, and Speech Transcripts Data. This dataset emphasizes human-verified quality across various stages of its curation.

Web Data. Our web data, constituting the majority of Sangraha, diverges from traditional Common Crawl-based approaches by prioritizing data quality. This involves manual verification of each website before scraping. Following Kakwani et al. (2020); Doddapaneni et al. (2023), we identify web sources, primarily news articles, through existing repositories and automated searches. Additionally, we engage volunteers to select websites in Indian languages and English based on content quality and cultural relevance. To enhance diversity, we manually verify and add a small subset of base URLs extracted from the mC4 corpus. The selection and scraping processes use the *webcorpus* toolkit[6]. We detail the process further in Appendix A.1.

PDF Data. Acknowledging the wealth of Indian language content in undigitized books and documents, we focus on text extraction from digitized PDFs. We download Indian language PDFs from the Internet Archive, selecting high-quality documents through a detailed process described in Appendix A.2. Additionally, we collect documents from different government sources including Parliamentary debates, magazines, textbooks, etc. We list down the sources and their details in Appendix A.2. For OCR, we employ GCP's Vision tool, recognized for good performance across categories (Dilmegani, 2023). Our future work will continue to explore digitization and OCR of new public sources.

Speech Transcripts. We source movie subtitles from OpenSubtitles[7], song lyrics, *Mann ki Baat* transcripts following Siripragada et al. (2020), and NPTEL transcripts[8], as extended by Bhogale et al. (2023b). These transcripts feature a substantial amount of technical text in Indian languages. Additionally, we transcribe around 80K hours of Hindi videos from YouTube using the Riva Conformer ASR Model[9]. We plan to extend the transcription efforts to all 22 scheduled Indian languages, with pipeline details described in Appendix A.3.

3.1.2 Sangraha Synthetic

There is a huge disparity between the information rich digital content and knowledge available in English as compared to Indian languages. To address this disparity, we introduce SANGRAHA SYNTHETIC, an initiative aimed at democratizing access to knowledge by translating a knowledge-rich English corpus into Indian languages. Utilizing INDICTRANS2 (Gala et al., 2023), we translated the entirety of English Wikimedia into 14 Indian languages resulting in nearly 90B tokens. Since INDICTRANS2 operates at the sentence-level and does not retain the document level formatting such as newlines, markdowns and other structures, we developed the SETU-TRANSLATE pipeline. This pipeline facilitates the translation of documents and conversations while preserving the original document structure.

Recognising the prevalent trend of "Romanized" Indic language usage particularly in informal settings and in digital communication, we extend Husain et al. (2024) and transliterate the above-translated content in 14 languages to Roman script using INDICXLIT (Madhani et al., 2023b) resulting in about 72B tokens. Going forward, we will extend SANGRAHA SYNTHETIC to cover all the 22 scheduled languages of India.

3.1.3 Sangraha Unverified

We introduce the SANGRAHA UNVERIFIED split to expand the Sangraha corpus while ensuring high quality. This split employs a perplexity filtering pipeline, inspired by CCNet (Wenzek et al., 2020), with the SANGRAHA VERIFIED split serving as the benchmark for data quality. Following this approach, we train 5-gram Kneser-Ney models using KenLM library (Heafield, 2011) for each language on a sample of SANGRAHA VERIFIED data. We then clean the entire Indic splits of CULTURAX (Nguyen et al., 2023a) and MADLAD-400 (Kudugunta et al., 2023) datasets through the Setu data cleaning pipeline. We consider CULTURAX and MADLAD-400 as these represent the latest and most comprehensive multilingual collections. We compute the perplexity for all the resultant cleaned documents and retain only those documents whose perplexity is below the threshold chosen for that language. We describe more details of the pipeline and the statistics for various languages in Appendix B.

3.2 Setu: A Comprehensive Pipeline for Data Cleaning, Filtering, and Deduplication

To clean, filter, and deduplicate Web, PDF, and Speech data, we create Setu, a pipeline built on Apache Spark which broadly has 4 stages - document preparation, document cleaning and analysis, flagging and filtering, and deduplication. The document preparation stage focuses on extracting the text from our diverse sources and creating text documents for further processing. For Web documents, we use *trafilatura* (Barbaresi, 2021b) to extract text from HTML, while the PDFs are run through a pipeline that uses a combination of the various bounding box related information to fil-

[6]https://github.com/AI4Bharat/webcorpus
[7]https://www.opensubtitles.org/
[8]https://nptel.ac.in/translation
[9]https://catalog.ngc.nvidia.com/orgs/nvidia/teams/riva/models/speechtotext_hi_in_conformer

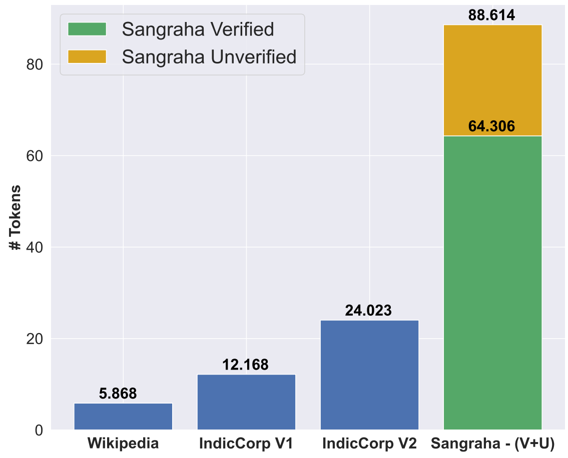


Figure 2: Comparison of the number of tokens (in Millions) in - INDICCORP V1, INDICCORP V2 and SANGRAHA VERIFIED + SANGRAHA UNVERIFIED

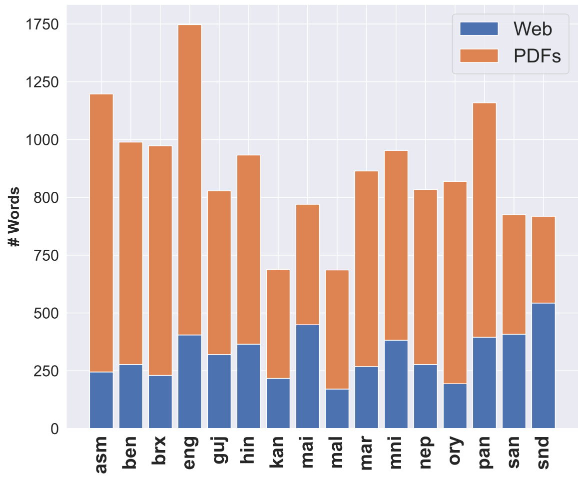


Figure 3: Average Document Size for Web and PDF documents in number of words.

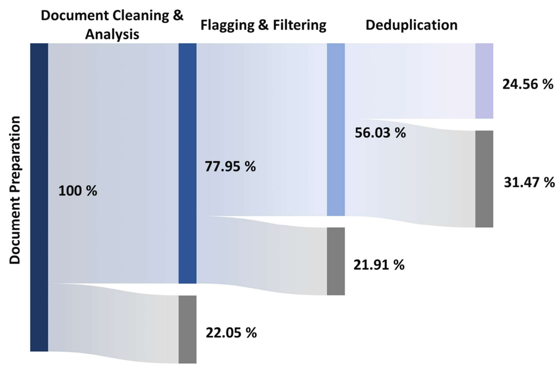


Figure 4: Percentage drop across the different stages of Setu when cleaned on SANGRAHA VERIFIED

ter out pages with potential recognition issues and noise. Text from speech data is extracted by either cleaning the SRT files to eliminate timestamps and other noise or by running the ASR transcripts through INDICPUNCT (Gupta et al., 2022) punctuation module to create text documents.

In the cleaning and analysis stage, we perform in-document cleaning to reduce the noise within a single document. We also use a multi-model approach for language identification by combining the outputs from INDICLID (Madhani et al., 2023a), CLD3, and NLLB (Costa-jussà et al., 2022). We then perform analysis by computing various statistics like character, word counts, NSFW word count, n-gram repetition ratio, etc. In the flagging and filtering stage, we apply various filters based on the statistics computed like line length filters, NSFW word filters, and repetition filters which remove noisy and toxic documents. In the end, the deduplication stage performs fuzzy deduplication using MinHashLSH implemented in *text-dedup*[10] repository by employing $n = 5$ and *threshold* $= 0.7$. A detailed overview and analysis of Setu is in Appendix D.

3.3 Data Analysis

The final statistics of SANGRAHA are shown in Table 1.

Comparison with other Multilingual Corpora: We compare SANGRAHA VERIFIED split with other Indic-only corpora - INDICCORP V1 (Kakwani et al., 2020), INDICCORP V2 (Doddapaneni et al., 2023) and Wikipedia. Figure 2 shows the distribution of the number of tokens for different Indic languages. We observe a significant increase in the size for all languages especially in the lower resource languages. Overall SANGRAHA VERIFIED contains 64.3B tokens and is 2.6× bigger than INDICCORP V2. We show a detailed language-wise comparison in Table 20 in Appendix F

Average document length comparison across languages: Figure 3 compares the average document length across various languages in terms of number of words. For Web Data, a single webpage is a document whereas for PDF data, a batch of consecutive pages is considered a document. We observe that Dravidian languages, i.e., Tamil, Malayalam, Kannada, and Telugu show considerably smaller document lengths, primarily because these languages are agglutinative.

How much data gets filtered by Setu? We present a comprehensive analysis of the attrition in token count observed across the various stages of

[10] https://github.com/ChenghaoMou/text-dedup

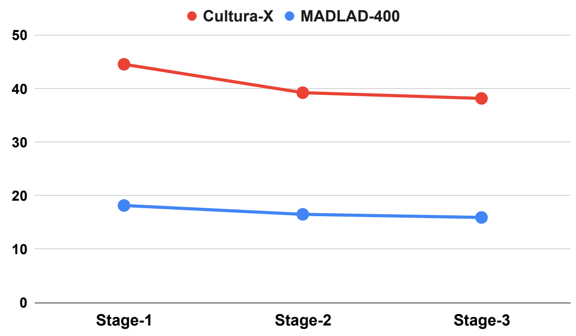


Figure 5: Number of tokens (in Billions) dropped at each stage in CULTURAX and MADLAD-400 when cleaned using Setu.

the Setu pipeline in Figure 4. Notably, the Deduplication stage exhibits the most significant reduction in tokens, which can be attributed to the fact that a lot of web content for Indic Languages comprises news articles with similar content disseminated across various platforms. We show qualitative examples of the content that gets filtered out at each stage in Appendix D. To show how Setu performs on other corpora, we clean the entire CULTURAX and MADLAD-400 datasets through Setu. Figure 5 shows the token drops across the stages for both. The massive drop from Stage-1 to Stage-2 shows that both the corpora had significant amount of noise inside documents like menus, headers, etc despite the claims of them being clean. We show examples of what kind of content is getting removed from CULTURAX and MADLAD-400 in Appendix C.

4 IndicAlign

INDICALIGN comprises two distinct splits: INDICALIGN - INSTRUCT and INDICALIGN - TOXIC data, each contributing to the robustness and diversity of the dataset. Table 2 encapsulates the overall statistics of INDICALIGN.

4.1 IndicAlign - Instruct

The INDICALIGN - INSTRUCT segment encompasses datasets that can be used to imbibe instruction-following ability in Large Language Models. Firstly we amalgamate different existing Instruction Finetuning (IFT) datasets with prompts authored by humans and responses generated by either humans or open, license-friendly models. To complement this human-centric approach, which is often too expensive and time consuming, we turn to synthetic data generation using existing chat-aligned models following the works of Ding et al. (2023b), Habash et al. (2022), and Xu et al. (2023b). We ensure that our outputs are always from open, license-friendly models and are always grounded in context. Given limited space, the descriptions given below are brief and we point the reader to Appendix E for more details and examples of all the datasets mentioned below.

Indic-ShareLlama We collect the prompts from the first turns of ShareGPT data[11] and prompt LLAMA2-70B CHAT model (Touvron et al., 2023b) for responses. Excluding non-English, coding, and math prompts, we translate and transliterate these prompt-response pairs into 14 languages.

Dolly-Translated Following Gala et al. (2024) and Husain et al. (2024), we translate and transliterate DOLLY-15K (Conover et al., 2023) dataset into 14 Indic languages.

OpenAssistant-Translated We extend the efforts of Gala et al. (2024); Husain et al. (2024) and release the translated and transliterated OPENASSISTANT (Köpf et al., 2023) in 14 Indic languages.

WikiHow Wikihow is an online wiki-style platform that serves as a valuable resource for a diverse array of how-to articles. Gala et al. (2024) curate around 20,400 and 6000 instruction-answer pairs in English and Hindi. The data is formulated as a completion task given either a question or a question along with a few initial steps. We translated it into 14 Indic languages.

IndoWordNet To get grammar and language creativity data we employ IndoWordnet (Bhattacharyya (2010), Panjwani et al. (2018)) to construct instruction-answer pairs. Our approach involves identifying a set of 21 distinct intents, such as part-of-speech identification and sentence creation using specific words. For each intent, we create five unique templates, both for prompts and responses, across 18 languages. Then for every WordNet entry, we randomly populate 20 templates, thus creating tailored instruction-answer pairs.

Anudesh Anudesh is a crowd-sourced collection of prompts accompanied by responses from LLAMA2-70B CHAT model. The participants are provided with instructions detailing the nature of interaction expected from them. Each instruction has an Intent, a domain, and a language instruction. The intent describes the interaction's goal, such as summarization or recommendation seeking. The domain spec-

[11] https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

Component	Prompt source	Response source	Original/ Translated	#Examples	Avg. Turns	Avg. Inst. Len	Avg. Out. Len	#Lang.	Lexical Diversity
Indic ShareLlama	Human	Model	T	21.1k	1	60.45	267.98	15	57.69
Dolly-T	Human	Human	T	15.0k	1	12.34	59.38	15	47.23
OpenAssistant-T	Human	Human	T	19.9k	2.98	25.72	136.37	15	59.75
WikiHow	Human	Human	O	26.4k	1	43.85	327.95	2	23.87
IndoWordNet	Human	Human	O	74,272.2k	1	19.74	14.84	18	37.24
Anudesh	Human	Model	T	43.3k	1.58	12.4	149.28	20	51.69
Wiki-Conv	Model	Model	T	144k	9.14	7.09	11.22	15	23.17
Wiki-Chat	Model	Model	T	202k	2.8	23	227.75	15	56.67
HH-RLHF-T	Human	Model	T	32.6k	1	14.11	64.88	15	79
Toxic Matrix	Model	Model	T	90.3k	1	33.68	89.64	15	86.57

Table 2: Overall statistics of INDICALIGN. Dolly-T represents Dolly Translated, OpenAssistant-T represents OpenAssistant Translated

ifies the context, like "Indian Festivals" or "Food and Cuisine," where the interaction needs to unfold. Language sets the linguistic framework, guiding participants to create prompts in designated languages, including - Indic languages, English, transliterated Indic, and a code-mix of Indic and English. Following these instructions, participants create prompts that are subsequently coupled with LLAMA2-70B CHAT model's responses, translated to match the initial language specification.

Wiki-Conv We use Wikipedia passages and Wiki-Infoboxes as contexts to generate conversations. An Infobox is a fixed-format table added to Wikipedia articles that summarizes important facts, statistics, and important points in an easy-to-read format. We use LLAMA2-70B CHAT (Touvron et al., 2023b) to generate an entire conversation in a user-assistant format in a single generation and subsequently translate and transliterate it to 14 Indic languages using our pipeline.

Wiki-Chat Unlike WIKI-CONV, here we try to simulate dialogues between two LLMs. We use the Wikipedia context from WIKI-CONV to determine an intent which drives the conversation between a User LLM agent and an Assistant LLM agent. This simulation involves four distinct LLM agents: Intent LLM, Init User LLM, Assistant LLM, and Next User LLM. We use LLAMA2-70B CHAT (Touvron et al., 2023b) and MIXTRAL-8X7B-V0.1 (Jiang et al., 2024b) to simulate the conversations which are then translated and transliterated to 14 Indic languages.

We again request the reader to check Appendix E for more details and examples (especially for IndoWordNet, Anudesh, Wiki-Conv and Wiki-Chat).

4.2 IndicAlign - Toxic

Aligning chat models to responsibly handle toxic prompts is a crucial aspect of developing ethically responsible models. In this work, we present initial steps towards creating datasets aimed at refining model responses to toxic inputs. We use both human and synthetic data collection strategies and introduce two distinct datasets: HH-RLHF-Translated, comprising human-curated data, and Toxic Matrix, a novel toxic alignment dataset created synthetically.

HH-RLHF - Translated We prompt LLAMA2-70B CHAT to classify each of the initial user prompts from HH-RLHF (Bai et al., 2022) as either toxic or non-toxic, along with providing the rationale for its decision. From approximately 169K initial prompts, around 32K were identified as toxic. We frame the response for each toxic prompt by including a statement of inability to engage due to the toxic nature of the prompt accompanied by the rationale given by the model. We then translate this into 14 Indic languages followed by transliteration to Roman script.

Toxic Matrix We introduce a novel approach to generate toxic alignment data synthetically using a taxonomy with three main axes: Content-Type, Target Group, and Prompt Style. We expand each axis to come up with a comprehensive list of categories. Table 3 shows examples of categories under each of the axes. We then use MISTRAL-7B CHAT (Jiang et al., 2023) model to generate toxic prompts for each combination of our categories. We found MISTRAL-7B CHAT to have miniamal safety alignment which allows it to create highly creative toxic content. We then use another model - LLAMA2-70B CHAT - to respond to these toxic prompts. LLAMA2-70B CHAT is selected for its strong toxic alignment, meaning it either refuses to engage with the toxic content or provides a non-toxic response. We generate in total about 90K toxic prompt-response pairs, all of which are translated and transliterated to 14 Indic languages.

Although previous works have shown different ways to distill instruction following alignment from strong models we propose this method as one of the ways to distill toxic alignment using a combination of a weakly and a strongly toxic-aligned model. This approach, while still under development, offers a promising direction for improving the ethical alignment of conversational models. It's important to note, however, that this method is part of an ongoing effort and not a definitive solution to ensuring toxic alignment. We propose this taxonomy-based approach as one of the ways of approaching this problem of generating/collecting toxic data and thereby aligning the models.

We refer the reader to Appendix F for further details and examples.

4.3 Analysis

Number of turns: Our curated dataset exhibits a wide range across various dimensions. Specifically, the range of dialogue turns spans from an average of 9.27 to a minimum of 1, which will result in the trained model's capability to support dialogues of both short and extended lengths. Furthermore, the variation in average instruction and output lengths will underscore the model's proficiency in processing and generating content of diverse lengths.

Lexical diversity of INDICALIGN data: To show the lexical diversity of the prompts, following the work of UltraChat (Ding et al., 2023b) we use the Measure of Textual Lexical Diversity (MTLD) score (McCarthy and Jarvis., 2010). As seen in Table 2, the OpenAssistant dataset has the highest lexical diversity, attributable to its sourcing from approximately 13,500 volunteers. Additionally, the lexical diversity of the Wiki-Chat dataset is on par with other human-generated datasets such as Indic ShareLlama and Dolly, indicating that our methodology of using intents to drive conversations, is effective in producing prompts with diversity comparable to those collected from human participants.

Intent Diversity Analysis: Figure 6 depicts the distribution of intents within the WIKI-CHAT dataset. Notably, since we have used Wikipedia as the context, we understandably see a majority of the interactions revolving around Information

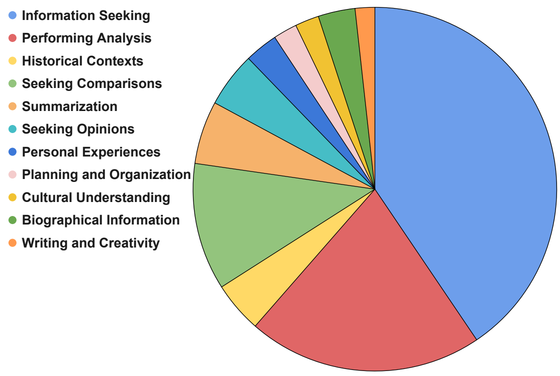


Figure 6: Wiki-Chat Intent Analysis - The different kinds of intents based on which Wiki-Chat conversations are simulated

Content Type	Fraudulent activities, Harassment on Accent, Vaccine Misinformation, Kidnapping, Harassment on Appearance, Ethnic Insults, Suicidal Ideation
Target Group	Children with Disabilities, Bengalis, Gujaratis, South Indians, Adolescents, Heterosexuals, Adults (30-49yrs), Sardarjis
Prompt Style	Direct, Indirect, Misleading, Long Con, Fooling, Provocative, Role-Play, Exploitative, Manipulative

Table 3: Examples for each axis in the Toxic Matrix taxonomy

seeking. We also observe diversity of intents centered around various real-world scenarios showing the real-world applicability of our data. We show additional analysis of the data in Appendix E.

5 Conclusion

In summary, our work addresses the underrepresentation of low and mid-resource languages, specifically focusing on the 22 constitutionally recognised languages. We introduce INDICLLMSUITE, a comprehensive framework encompassing SANGRAHA pretraining data, SETU a Spark-based pipeline for data curation, INDICALIGN - INSTRUCT a diverse prompt-response collection, and INDICALIGN - TOXIC containing aligned toxic responses for Indic LLMs. By striking a balance between human-verified content and model-generated data, we aim to provide equitable access to information for diverse linguistic communities. We encourage community collaboration in the costly endeavor of LLM training, advocating for the pooling of resources to build high-quality fully open source Indic language LLMs. Through the public release of our tools and datasets, we hope to inspire advancements in LLM development for Indian languages and beyond.

Limitations

While Sangraha leverages publicly available web content, PDFs, and videos as primary data sources, it's crucial to acknowledge potential biases inherent in this data, which could be inherited by any model trained on the data. We leave the analysis on potential biases and debiasing techniques for future work. We rely on NSFW word detection for toxic data detection, which does not fully capture or mitigate toxicity and sometimes results in false positives. We call upon the community to create better toxic data detection techniques for all Indian languages. Despite our efforts to remove Personally Identifiable Information (PII) from crowdsourced data, there remains a risk of inadvertent inclusion. The dataset exhibits lower representation from higher age groups, uneven coverage across Indian states, and a lack of comprehensive inclusion for low-resource languages. Additionally, our method of translating toxic prompts into Indic languages may not adequately capture the nuanced variations in Indian contexts.

We again call upon the community to contribute towards enhancing data diversity, improving translation methodologies for better cultural and contextual relevance, and developing more effective tools for debiasing and ensuring ethical use.

Ethics Statement

All individuals involved in this effort, including annotators and developers, were adequately compensated for their work, adhering to all relevant norms and regulations of our country. The volunteers engaged in the curation of crowd-sourced data were informed about the public release of the data. Both the pretraining and fine-tuning datasets have been checked for offensive content, as necessary.

The released code will carry an MIT License[12], and all datasets will be released under appropriate open licenses.

Acknowledgements

We would like to thank EkStep Foundation and Nilekani Philanthropies for their generous grant towards building datasets, models, tools and other resources for Indian languages. We also thank Google for their valuable grant that facilitated OCR through Google Cloud Vision and for access to free TPUs as part of the TPU Research Cloud (TRC) initiative. We thank the Sarvam AI team for their insightful discussions and constructive comments on this project. We are also immensely grateful to the volunteers from the AI4Bharat team for their motivation and meticulous efforts in conducting manual audits.

[12] https://opensource.org/licenses/MIT

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *CoRR*, abs/2303.09540.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Adrien Barbaresi. 2021a. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 122–131. Association for Computational Linguistics.

Adrien Barbaresi. 2021b. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2010. Indowordnet. In *Lexical Resources Engineering Conference 2010 (LREC 2010)*, Malta.

Kaushal Bhogale, Deovrat Mehendale, and Tahir Javed. 2024. Mahadhwani. `https://github.com/AI4Bharat/Mahadhwani`.

Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023a. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kaushal Bhogale, Sairam Sundaresan, A. Raman, Tahir Javed, Mitesh M. Khapra, and Pratyush Kumar. 2023b. Vistaar: Diverse benchmarks and training sets for indian language asr. *INTERSPEECH*.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Compression and Complexity of SEQUENCES 1997, Positano, Amalfitan Coast, Salerno, Italy, June 11-13, 1997, Proceedings*, pages 21–29. IEEE.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 380–388. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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, Loïc 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. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Cem Dilmegani. 2023. Ocr in 2024: Benchmarking text extraction/capture accuracy.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023a. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023b. Enhancing chat language models by scaling high-quality instructional conversations.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

István Endrédy and Attila Novák. 2013. More effective boilerplate removal - the goldminer algorithm. *Polibits*, 48:79–83.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned llm.

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. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Anirudh Gupta, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, Priyanshi Shah, Harveen Singh Chadha, and Vivek Raghavan. 2022. indic-punct: An automatic punctuation restoration and inverse text normalization framework for indic languages.

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022. Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2672–2681, Marseille, France. European Language Resources Association.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. 2022. Scaling laws and interpretability of learning from repeated data. *CoRR*, abs/2205.10487.

Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models models via romanization.

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. 2023. Mistral 7b.

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. 2024a. Mixtral of experts. *CoRR*, abs/2401.04088.

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. 2024b. Mixtral of experts.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. *CoRR*, abs/2304.07327.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Anoop Kunchukuttan. 2020. The indicnlp library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

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. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit.

Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023a. Bhasa-abhijnaanam: Native-script and romanized language identification for 22 Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. Association for Computational Linguistics.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. 2023b. Aksharantar: Open Indic-language transliteration datasets and models for the next billion users. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 40–57, Singapore. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocdd, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet*, Mysore, India.

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

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023b. Seallms – large language models for southeast asia.

Ritesh Panjwani, Diptesh Kanojia, and Pushpak Bhattacharyya. 2018. pyiwn: A python based API to access Indian language WordNets. In *Proceedings of* the 9th Global Wordnet Conference, pages 378–383, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116.

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. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. 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 M. 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 J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Scrapinghub. 2021. Article extraction benchmark. GitHub repository.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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. 2023a. Llama: Open and efficient foundation language models.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv: 2307.06018*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023b. Wizardlm: Empowering large language models to follow complex instructions.

Canwen Xu, Daya Guo, Nan Duan, and Julian J. McAuley. 2023c. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6268–6278. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM*, 59(11):56–65.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. (inthe)wildchat: 570k chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Domain	Number of Websites
com	5926
in	817
org	446
net	250
co.in	81
tv	75
others	445
Total	**8040**

Table 4: Sangraha Verified Website domain statistics

PDF Sources	#PDFs	#Pages
Internet Archive	437225	74M
eGyanKosh	5133	88K
Indian Parliament	30964	2.7M
AIR News	74353	148K
Govt. Magazines	895	46K
School Books	4315	359K
Miscellaneous	27988	4.6M
Total	**507419**	**82M**

Table 5: Sangraha PDF sources - The final statistic of the PDFs on which OCR has been performed.

Appendix

A SANGRAHA VERIFIED - Curation Details

A.1 Curation of SANGRAHA VERIFIED- Web Data

Here we discuss the details about curation of SANGRAHA VERIFIED - Web data. Table 4 shows the domain-level statistics of the websites in SANGRAHA VERIFIED.

In this work, we adopt a three-fold strategy to collect a comprehensive collection of websites for scraping. Firstly, we extend the efforts of Kakwani et al. (2020) and Doddapaneni et al. (2023) of discovering web sources using existing repositories and automated web searches, to discover a large list of Indic language websites. But, unlike the previous efforts, we do not restrict ourselves to just news websites. Secondly, we identify various domains such as Indian Culture, Food, Health, and Travel, among others and enlist volunteers to gather websites within these domains, prioritizing those in Indic languages or English but pertinent to the Indian context. Thirdly, we collect the base URLs from MC4 (Xue et al., 2021), focusing on websites with high amount of content and get them verified by volunteers. Additionally, we include all the Indian Government websites[13] which serves as a valuable resource, given their multilingual content.

Volunteers review each website collected via automated methods and decide on acceptance or rejection based on the criteria defined. A website can be rejected if either of the below conditions were met:

- Website is non-Indic or non-English.
- Website is an adult, gambling, or a general toxic website.
- Website has content that directly appears to be machine-translated.

Figure 7 presents the verification outcomes, highlighting a significant rejection rate due to website inactivity, particularly those sourced from MC4. This means that the information in existing collections if becoming outdated because of defunct websites. We make available the verification portal for further research utilization.

A.2 Curation of SANGRAHA VERIFIED - PDF Data

In this section, we elaborate on the methodology adopted for curating the SANGRAHA VERIFIED - PDF data. Table 5 shows the detailed source-level PDF statistics. We discuss the details of the curation of data from each source below.

Internet Archive

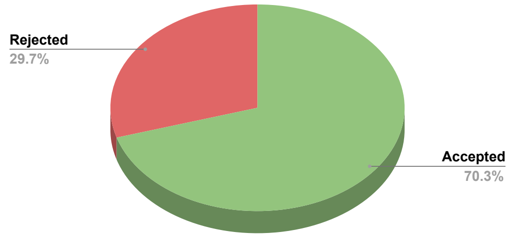
Utilizing the official API of the Internet Archive[14], we collected approximately 921K PDF documents across all Indic languages. This collection spans diverse categories such as religious texts, news articles, fiction, educational materials, and scientific literature. We subsequently filtered out PDFs incompatible with GCP Vision[15], specifically excluding languages like Bodo, Dogri, Kashmiri, Konkani, Maithili, Manipuri, and Sindhi due to their low-resource status, with plans for future inclusion.

To optimize for quality and manage costs, OCR was performed solely on high-quality PDFs. We

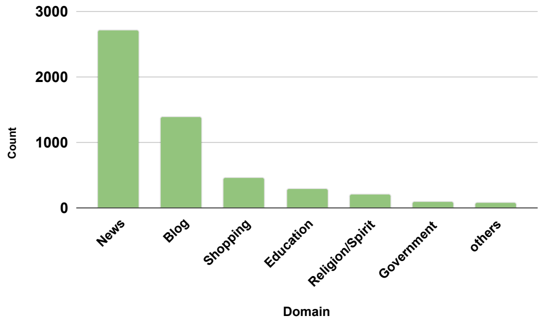
[13] https://igod.gov.in/

[14] https://archive.org/developers/internetarchive/

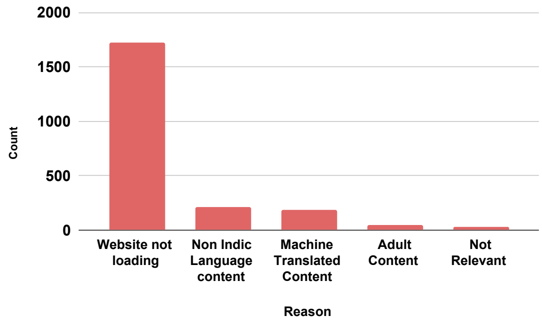
[15] https://cloud.google.com/vision/docs/languages



(a) Comparison of the number of websites accepted and rejected



(b) Accepted Websites statistics - Domain information



(c) Rejected Website statistics - Rejection Reason

Figure 7

first remove all the corrupted and encrypted PDFs. Additionally, resource limitations from GCP Vision necessitated the filtering of PDFs exceeding 2000 pages. We also filter out all the PDFs having less than 25 pages as these are often incoherent documents such as glossaries, comics, bills, and receipts. Quality assurance measures for OCR included filtering out scanned PDFs with a Pixel Per Inch (PPI) rating below 300. Additionally, we analyzed images from 10 consecutive pages of each PDF, focusing on metrics like average image area coverage and brightness. PDFs with images were considered for further analysis if they covered less than 50% of the page area and had a brightness level above 200. Table 6 shows the statistics of PDFs filtered after each filtering stage.

eGyanKosh

eGyanKosh, India's National Digital Repository, serves as a repository for digital learning resources from Open and Distance Learning Institutions, covering subjects such as History, Economics, Political Science, Public Administration, and Sociology, across various Indian languages.

Indian Parliament

This source comprises manually compiled summaries of debates and discussions from the Indian Parliament and various State Legislative Assemblies. These form a rich source of local and culturally relevant data. We collect all the publicly available Parliamentary and State Assembly materials. Table 7 shows the statistics of the state-wise collected documents.

AIR News

All India Radio (AIR) is the national radio broadcaster of India, a Prasar Bharati division, that streams radio programs in all major Indian languages. Following the approach of (Bhogale et al., 2023a), we collect news bulletins for 12 Indian languages. Table 8 shows the language level statistics of the collected data.

Govt. Magazines

We aggregated content from magazines published by governmental agencies, which include annual reports, details on governmental schemes, initiatives, cabinet decisions, and current affairs, published in multiple Indian languages.

Language	Original Count	After Validity Check	After Page Count Check	After Image Filters
Hindi	349,365	344,454	106,112	102,164
Urdu	177,867	157,121	127,495	73,966
Sanskrit	88,238	84,804	76,401	70,663
Bengali	59,636	55,023	50,825	45,272
Tamil	52,199	49,924	37,243	29,755
Telugu	50,320	48,919	40,860	38,243
Gujarati	43,677	42,021	34,514	34,038
Malayalam	34,858	31,594	11,627	4,725
Kannada	24,446	23,589	18,661	17,493
Punjabi	13,898	12,932	7,397	5,617
Marathi	9,710	9,174	7,875	7,478
Assamese	2,424	2,408	2,205	2,408
Nepali	1,545	1,497	836	671
Odia	4,972	4,733	2,439	4,732

Table 6: Statistics of PDFs filtering from Internet Archive

State	Number of PDFs
Andhra Pradesh	3383
Bihar	306
Gujarat	3241
Haryana	433
Himachal Pradesh	2035
Jharkhand	124
Karnataka	8405
Kerela	2039
Madhya Pradesh	656
Maharashtra	544
Punjab	287
Rajasthan	7
Tamil Nadu	680
Indian Parliament	14896
Total	**37036**

Table 7: Statistics of the PDFs collected from Indian Parliament

Language	Number of PDFs
Bengali	5721
Gujarati	5586
Hindi	18560
Kannada	4888
Konkani	471
Malayalam	5665
Marathi	8958
Nepali	1686
Odia	5769
Punjabi	885
Sanskrit	730
Tamil	7002
Telugu	5555
Urdu	2877
Total	**74353**

Table 8: Language-wise statistics of PDFs collected from AIR newsonair

State	Number of PDFs
Andhra Pradesh	126
Assam	61
Bihar	426
Goa	31
Haryana	31
Himachal Pradesh	1909
Karnataka	502
Kerala	121
Maharashtra	76
Manipur	70
Meghalaya	293
Mizoram	40
Nagaland	681
Odisha	41
Punjab	195
Rajasthan	186
Telangana	235
Tripura	365
West Bengal	125
National	598
Other Books	1442
Total	**7554**

Table 9: State-wise statistics of School textbooks collected

School Textbooks

This set includes publicly available textbooks from various Indian states and those published by the National Council of Educational Research and Training (NCERT), providing a rich source of educational content in multiple Indian languages. Table 9 shows the statistics of the books collected from different sources.

Miscellaneous

In addition to the categorized sources, we also incorporated a variety of documents from government and public domains, focusing on content either in Indic languages or in English with relevance to India.

A.3 Curation of SANGRAHA VERIFIED - Speech Data

Here we discuss the details about Speech Data split of SANGRAHA VERIFIED component. Table 10 shows the detailed source-level statistics. We discuss the details of the curation of the data from each source below.

Source	Number of Instances
YouTube - Hindi	276K videos
Open Subtitles	14K movies
NPTEL - Transcripts	1.4K courses
Mann Ki Baat	1.4K podcasts
Others	15K
Total	**309K**

Table 10: Statistics of the various sources of Speech Data collected

Youtube - Hindi

Following the approach of Bhogale et al. (2024), we collect around 80K hours of audio data from Youtube videos in Hindi language. We then chunk it into smaller segments by detecting silences using WebRTC VAD[16] and get each chunk transcribed using the Hindi Conformer model. Then, we piece together all the transcripts to obtain the transcript for the whole video.

OpenSubtitles

Following Gao et al. (2021), we collect all the Indic Language subtitles from OpenSubtitles[17]. We first process the SRT files using simple regex based patterns to remove the timestamps and extract the text. We then define regex patterns to filter out other noisy content like character cues, continuation ellipses, etc. We then combine the different parts to form a single document per SRT file. Table 11 shows the language-wise statistics of Subtitles.

NPTEL - Transcripts

The National Programme on Technology Enhanced Learning (NPTEL)[18] is an Indian e-learning platform for university-level science, technology, engineering and mathematics subjects that is jointly developed by various Indian Institutes. Although the course content developed by NPTEL is primarily in English, a lot of it has been manually transcribed and translated into 11 different Indian Languages and reviewed before being made publicly available. The translated content has been compiled and released as course textbooks. Table 12 shows the statistics of the course transcripts available in different languages.

[16] https://github.com/wiseman/py-webrtcvad
[17] https://www.opensubtitles.org/
[18] https://nptel.ac.in/

Language	Number of Instances
Assamese	2
Bengali	2619
English	1178
Hindi	2808
Kannada	7
Malayalam	7571
Odia	3
Sindhi	30
Tamil	223
Telugu	20
Urdu	129
Total	**14590**

Table 11: Language wise statistics of subtitles collected from OpenSubtitles

Language	Number of Courses
Assamese	1
Bengali	91
English	523
Gujarati	106
Hindi	184
Kannada	89
Malayalam	108
Marathi	85
Punjabi	1
Tamil	150
Telugu	98
Total	**1436**

Table 12: Language wise statistics of the course transcripts collected from NPTEL

Language	Number of Instances
Assamese	63
Bengali	91
English	410
Gujarati	92
Hindi	89
Kannada	78
Malayalam	89
Marathi	90
Manipuri	65
Odia	82
Punjabi	81
Tamil	85
Telugu	89
Urdu	64
Total	**1468**

Table 13: Language-wise Mann Ki Baat transcripts collected

Mann Ki Baat

Mann Ki Baat is an Indian Radio programme hosted by the Indian Prime Minister usually with a frequency of 1 per month. This is transcribed and then manually translated into 13 Indian languages. Table 13 shows the language wise statstics.

A.4 Setu Translate

Majority of the machine translation systems are trained as sentence-level translators which often struggle to preserve various entities like inter-sentence separators, new-line characters, tab-spaces, markdowns, bullet points, etc. Simple sentence-tokenizers present in the packages like NLTK (Loper and Bird, 2002) and IndicNLP Library (Kunchukuttan, 2020) are not capable of retaining these inter-sentence separators and markdowns. We introduce SETU-TRANSLATE, a robust translation pipeline for mass-translation of both pre-training as well as Instruction fine-tuning data while preserving the structure of the document. Overall, SETU-TRANSLATE focuses on the accurate identification of the parts of the document that must be sent to the translation model and then the replacement of the translated sentences in the overall document thereby preserving the overall structure of the translated document. The five main of SETU-TRANSLATE are described in this section.

Templating

Using regex patterns, we identify the parts of the documents we intend to translate. The goal of this stage is to preserve the structure of the document. The regex patterns defined ignore markdown structures, code snippets (enclosed in backticks), bullet points, paragraph indicators, Roman numerals, etc., and extract only the sentences. After performing unicode-normalization and deduplication on the extracted sentences, a global sentence-level dataset is created.

Inference

We binarize the data first and then utilize INDICTRANS2 for translating English into Indic languages. We leverage both GPUs and TPUs for large-scale translation. To benefit the community, we open-source the flax port for INDICTRANS2 for faster TPU inference.

Replace

Once we have the translated sentences, we perform a regex-based replacement of the original sentences with the translated ones. This ensures that only sentences are replaced and the other structure of the document is retained as is.

A.5 Setu Transliteration

Similar to translation, we also release the Setu Transliteration pipeline. Since transliteration is done at a word level and doesn't consider the context of the remaining words, we follow the normal word replacement strategy. We maintain a continuously updating mapping of Indic words to their Roman counterparts in a prefix-based hierarchical format which we feel is the key to speedup and rapid access to the required word pairs.

Word Mapping Dictionary

For the creation of the initial mapping, we use AKSHARANTAR(Madhani et al., 2023b) dataset, which is the largest publicly available transliteration dataset for Indic languages as the starting point. We convert AKSHARANTAR into the said prefix-based hierarchical format. This mapping is continuously updated with the new mappings as we discover new un-romanized words further in our pipeline.

Word Replacement

Word level replacement has 2 main challenges: (i) identifying words to replace while preserving the entire document structure; and (ii) unordered replacement leading to sub-word replacement instead of the entire word.. We address (i) using the same regex-based approach used in SETU-TRANSLATE. To address (ii), we sort the mapping based on source-language word length in descending order before feeding the mapping to the regex-based 'replace' module.

Inference

During the first 'replace' pass, we log the unromanized words whose mapping is not available in the current word mapping dictionary. In the 'inference' stage, we transliterate these words using INDICXLIT(Madhani et al., 2023b) to get an updated word-mapping dictionary. We then repeat the word-replacement until all the words are properly romanized.

B Curation of SANGRAHA UNVERIFIED

We describe the methodology employed for curating the SANGRAHA UNVERIFIED split, leveraging a perplexity-based filtering pipeline inspired by CCNET (Wenzek et al., 2020).

We first randomly sample 200,000 documents from SANGRAHA VERIFIED split for each language. We then normalize each document by converting text to lowercase, removing accents from characters, normalizing numbers to a uniform representation (specifically converting all digits to "0"), replacing a predefined set of Unicode punctuation with their ASCII counterparts, and removing non-printing characters. We then train a sentencepiece tokenizer and tokenize all of the sampled data. Then, we train a 5-gram Kneser-Ney models using KenLM (Heafield, 2011) library. We binarize these models for quicker inference.

For deciding the language-specific thresholds, we create a validation set by sampling another 100,000 documents from SANGRAHA VERIFIED and calculate the perplexity of each document using the trained n-gram models. We then sort the perplexities and choose the 80th percentile value as the threshold for each language. Table 14 shows the thresholds chosen for each language. To prefer more quality over volume, higher percentile thresholds can be chosen, but that may result in reduced diversity and representativeness of the resultant data.

We clean the entire CULTURAX and MADLAD-400 corpora using the Setu Cleaning pipeline and

Language	Min PP	Max PP	Mean PP	PP Threshlold	Total Docs	Chosen Docs	Filtering rate
Assamese	27.4	65155.6	1013.9	1216	25617	18713	26.9%
Bengali	6.7	22941.5	286.6	606.7	6838196	6274727	8.24%
Gujarati	7.8	23184.4	421.7	792.5	640843	586977	8.4%
Hindi	5.7	160264.7	230.44	378.8	19362407	17271194	10.8%
Kannada	8.6	25413.1	74.5	103.4	748914	623662	9.1%
Malayalam	5.6	43419.9	65.8	61.4	1723524	1012425	41.25%
Marathi	8.3	16032.2	214.2	277.8	1322324	1051722	20.4%
Nepali	7.1	20334.8	140.0	120.32	1625754	961637	40.84%
Odia	5.8	166311	160.0	170.8	61692	44298	28.1%
Punjabi	8.0	23375.0	232.6	229.7	302421	195115	35.48%
Sanskrit	32.8	5919.0	823.8	1397.7	3332	2993	10.17%
Tamil	6.2	22583.3	157.6	262.3	2416008	2089674	13.5%
Telugu	12.6	65297.8	139.3	377	930407	898991	3.37%
Urdu	2.4	25206.5	158.4	316.8	1502769	1372703	8.65%

Table 14: Perplexity Statistics of CULTURAX and MADLAD-400 datasets. Perplexity is calculated using n-gram language models trained on data sampled from SANGRAHA VERIFIED.

de-duplicate it with the entire SANGRAHA VERIFIED split. Finally, we calculate the perplexities of each document and filter out those that are above the chosen threshold. Table 14 shows the final number of documents chosen after perplexity filtering.

B.1 Perplexity Analysis

Figure 8 shows the perplexity distributions of the cleaned CULTURAX and MADLAD-400 data using the n-gram language models trained on SANGRAHA VERIFIED. we observe that certain languages, specifically Hindi, Malayalam, and Marathi, exhibit relatively tight distributions of perplexity values. This indicates a higher degree of similarity in the statistical properties of these language datasets to the SANGRAHA VERIFIED training data. Conversely, we note that some languages, particularly those classified as low-medium resource, show more dispersed perplexity distributions.

C Uncleanliness of Existing Corpora

Issues with Language Identification

The evolution of Language Identification (LID) models has predominantly focused on European languages, leading to significant challenges in accurately identifying languages from diverse linguistic families, notably Indic languages. Kreutzer et al. (2022) highlights a significant concern regarding the mislabeling of languages in existing multilingual corpora, an issue that undermines the reliability of language identification (LID) models. In this small study, we analyze 200,000 documents per Indic language from the MC4 (Raffel et al., 2020) and OSCAR (Abadji et al., 2022) datasets, employing the INDICLID model for its superior performance on Indic languages and support for Romanized text (Madhani et al., 2023a). MC4 uses only *cld3* model whereas OSCAR defines an even stricter pipeline for identifying the language. It combines sentence-level LID and aggregates them based on certain thresholds to classify a document as multilingual or monolingual.

Our analysis uncovers a significant discrepancy in the accuracy of LID across various Indic languages within the MC4 dataset. The languages sharing a common script, such as Hindi, Marathi, and Nepali, experience higher rates of mislabeling. This contrasts with languages with unique scripts showing significantly lower mismatch percentages.

Conversely, the application of a more sophisticated LID methodology in the OSCAR dataset markedly diminishes these inaccuracies, showing the effectiveness of a refined approach to language identification. This observation demonstrates the necessity for the development of language family-specific identification models (Madhani et al., 2023a), as well as the incorporation of better LID modules within data-cleaning pipelines.

Amount of Noise in Existing Corpora

We clean the entirety of CULTURAX and MADLAD-400 datasets using our Setu cleaning pipeline and show the drop in the number of words and documents across the stages. This helps us identify the type of noise present in these datasets. Figure 5 shows the drop in the number of tokens in these datasets respectively. We see a significant drop in both from Stage-1 to Stage-2 showing that a lot of noise in the form of Menu Items, Index

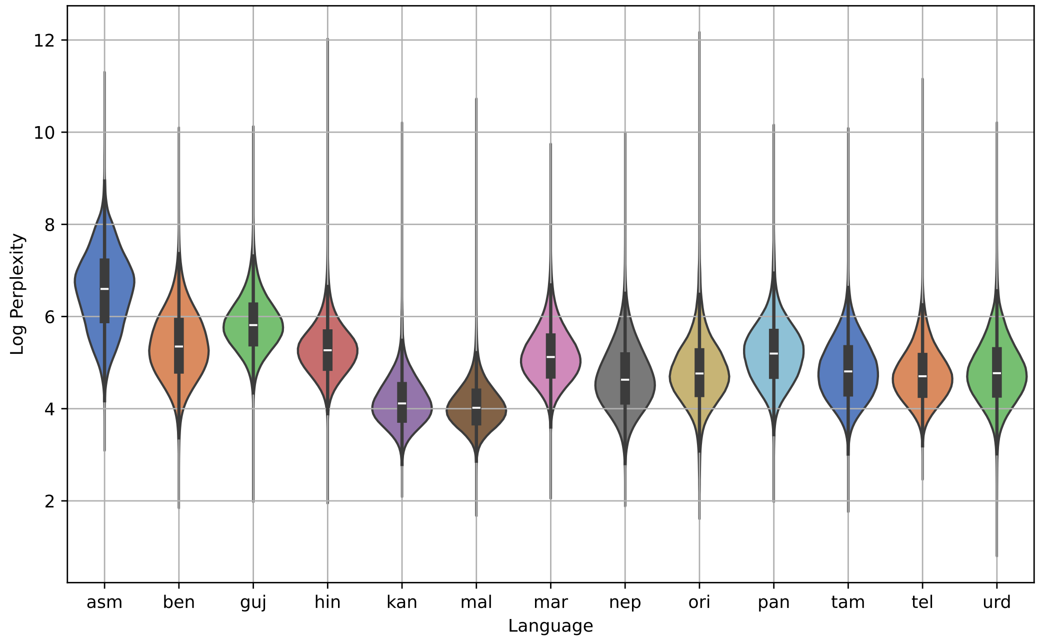


Figure 8: Log Perplexity distributions of Cleaned CULTURAX and MADLAD-400 using 5-gram language models trained on SANGRAHA VERIFIED

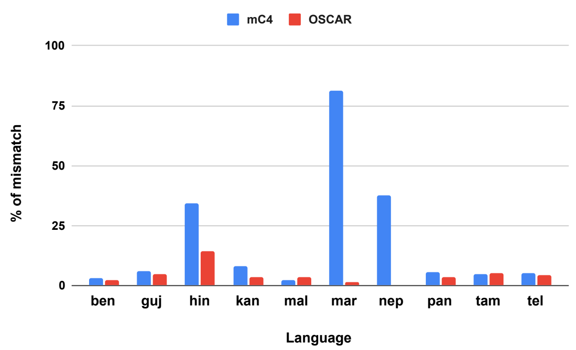


Figure 9: % mismatch of the tagged language and the language predicted by INDICLID

lists, etc. must have crept in despite they being cleaned using their existing cleaning pipelines. We show a few examples of the kind of noisy text being filtered out in Figure 10. Table 21 shows the overall statistics of the CULTURAX data filtered out at each stage in Setu.

D Setu Data Cleaning Pipeline

Here we discuss the inner details for each stage in the SETU. Our main goal for SETU is to open-source a distributed and cloud-agnostic data cleaning pipeline for large-scale datasets so that community is not stuck with any specific cloud provider or compute-scale. Using spark, we are able to achieve all the necessary requirements. Figure 11 shows the overview of the entire pipeline.

D.1 Stage - 1: Document Preparation

This stage focuses on the extraction of text from varied data sources, ensuring the retention of main content while eliminating extraneous information and then preparing the notion of a document that is preserved throughout the pipeline. Due to the different modalities of content, this stage is different for each of Web, PDF, and Speech data.

Web Documents

Preparation of the document for Web data is quite straightforward. We use *trafilatura* (Barbaresi, 2021a) to extract the text from the HTML pages that are scraped by *webcorpus* scraper. Although *trafilatura* is reportedly the best non-commercial library (Scrapinghub, 2021), we still notice a considerable amount of noise in the outputs, specifically in dynamic webpages. Figure 12 shows an example of noisy content extracted using *trafilatura*. In Web data, each webpage after text extraction is considered as a document.

Cultura-X: Uncleaned

ભરૂચ : 5 અને 10 રૂપિયા નહિ સ્વીકારનારા વેપારીઓ સામે નોંધાશે રાજદ્રોહનો ગુનો | bharuch shopkeepers if not accept 5 and 10 rs case will be registered
Home > ગુજરાત > ભરૂચ : 5 અને 10 રૂપિયા નહિ સ્વીકારનારા વેપારીઓ સામે નોંધાશે રાજદ્રોહનો ગુનો
ભરૂચ : 5 અને 10 રૂપિયા નહિ સ્વીકારનારા વેપારીઓ સામે નોંધાશે રાજદ્રોહનો ગુનો
સિકકાઓ સ્વીકારવામાં નહિ આવતાં હોવાની ફરિયાદો, ભરૂચના અધિક જિલ્લા મેજીસ્ટ્રેટે બહાર પાડયું જાહેરનામુ.
BY Connect Gujarat5 July 2021 10:31 AM GMT
Connect Gujarat5 July 2021 10:31 AM GMT
ભરૂચના અધિક જિલ્લા મેજીસ્ટ્રેટે બહાર પાડેલા એક જાહેરનામાના કારણે વેપારીઓ અને રીકશાચાલકો સહિત અન્ય વ્યવસાયકારોમાં ફફડાટ ફેલાયેલો છે. જે લોકો 5 અને 10 રૂપિયાના ચલણી સિકકાઓ સ્વીકારતાં નથી તેમની સામે રાજદ્રોહનો કેસ કરવાની ચીમકી આપવામાં આવી છે.
ભરૂચ જિલ્લાના ગામડાઓ અને શહેરી વિસ્તારમાં ભારતીય રીઝર્વ બેન્ક RBI દ્વારા બહાર પાડેલા ચલણી નોટ તથા સિક્કા નહિ ચલાવનારાને ટકોર કરાઈ છે. જિલ્લાના ગામડાઓ તથા શહેરી વિસ્તારમાં ભારતીય રીઝર્વ બેન્ક દ્વારા બહાર પાડેલ ચલણી નોટો તથા સિક્કા ચલાવવામાં આવતા નથી કે સ્વીકારવામાં આવતાં નથી તેવી ફરિયાદો તંત્રના ધ્યાને આવી હતી. ખાસ કરીને ₹ 5 ની ચલણી નોટ તથા 10 ના સિક્કાઓ ભારતીય રીઝર્વ બેંક દ્વારા માન્ય કરેલા હોવા છતાં સ્વીકારવાની વેપારીઓ અને લોકો આનાકાની કરતા હોવાના કિસ્સાઓ બહાર આવી રહ્યાં હતાં.જેને લઈ અધિક જિલ્લા મેજીસ્ટ્રેટ જે.ડી.પટેલે એક જાહેરનામુ બહાર પાડયું છે.
ચલણમાં રહેલા નોટો કે સિક્કાઓનો અસ્વીકાર કરનાર વ્યક્તિ સામે રાજદ્રોહનો ગુનો નોંધાઇ શકે છે. હાલ 10 ની ચલણી નોટો જૂની અને ઓછા પ્રમાણમાં બજારમાં વ્યવહારમાં હોવા સામે 10 ના સિક્કા વધુ હોવાથી તે ફરી રહ્યાં છે.કેટલાક વેપારીઓને ચિલ્લર અને સિક્કાઓનો ભાર વધુ લાગતો હોવાથી તેઓ જાતે જ તેમની પાસે થતા ભરાવાને લઈ તેનો લોકો પાસેથી સ્વીકાર કરવાની ચોખ્ખી ના પાડી દે છે. જોકે ભારતીય ચલણમાં રહેલા કોઈપણ સિક્કા કે નોટોનો અસ્વીકાર કોઈ ફરિયાદ કરે તો આવા લોકો કે વેપારીને રાજદ્રોહના ગુનાનો ભોગ બનાવી શકે છે.

Cultura-X: cleaned

સિકકાઓ સ્વીકારવામાં નહિ આવતાં હોવાની ફરિયાદો, ભરૂચના અધિક જિલ્લા મેજીસ્ટ્રેટે બહાર પાડયું જાહેરનામુ.

ભરૂચના અધિક જિલ્લા મેજીસ્ટ્રેટે બહાર પાડેલા એક જાહેરનામાના કારણે વેપારીઓ અને રીકશાચાલકો સહિત અન્ય વ્યવસાયકારોમાં ફફડાટ ફેલાયેલો છે. જે લોકો 5 અને 10 રૂપિયાના ચલણી સિકકાઓ સ્વીકારતાં નથી તેમની સામે રાજદ્રોહનો કેસ કરવાની ચીમકી આપવામાં આવી છે.
ભરૂચ જિલ્લાના ગામડાઓ અને શહેરી વિસ્તારમાં ભારતીય રીઝર્વ બેન્ક RBI દ્વારા બહાર પાડેલા ચલણી નોટ તથા સિક્કા નહિ ચલાવનારાને ટકોર કરાઈ છે. જિલ્લાના ગામડાઓ તથા શહેરી વિસ્તારમાં ભારતીય રીઝર્વ બેન્ક દ્વારા બહાર પાડેલ ચલણી નોટો તથા સિક્કા ચલાવવામાં આવતા નથી કે સ્વીકારવામાં આવતાં નથી તેવી ફરિયાદો તંત્રના ધ્યાને આવી હતી. ખાસ કરીને ₹ 5 ની ચલણી નોટ તથા 10 ના સિક્કાઓ ભારતીય રીઝર્વ બેંક દ્વારા માન્ય કરેલા હોવા છતાં સ્વીકારવાની વેપારીઓ અને લોકો આનાકાની કરતા હોવાના કિસ્સાઓ બહાર આવી રહ્યાં હતાં.જેને લઈ અધિક જિલ્લા મેજીસ્ટ્રેટ જે.ડી.પટેલે એક જાહેરનામુ બહાર પાડયું છે.
ચલણમાં રહેલા નોટો કે સિક્કાઓનો અસ્વીકાર કરનાર વ્યક્તિ સામે રાજદ્રોહનો ગુનો નોંધાઇ શકે છે. હાલ 10 ની ચલણી નોટો જૂની અને ઓછા પ્રમાણમાં બજારમાં વ્યવહારમાં હોવા સામે 10 ના સિક્કા વધુ હોવાથી તે ફરી રહ્યાં છે.કેટલાક વેપારીઓને ચિલ્લર અને સિક્કાઓનો ભાર વધુ લાગતો હોવાથી તેઓ જાતે જ તેમની પાસે થતા ભરાવાને લઈ તેનો લોકો પાસેથી સ્વીકાર કરવાની ચોખ્ખી ના પાડી દે છે. જોકે ભારતીય ચલણમાં રહેલા કોઈપણ સિક્કા કે નોટોનો અસ્વીકાર કોઈ ફરિયાદ કરે તો આવા લોકો કે વેપારીને રાજદ્રોહના ગુનાનો ભોગ બનાવી શકે છે.

(a) CULTURAX

MADLAD: Uncleaned

করোনার টিকা নিলেন মাননীয় মেয়র লিটন ও পরিবারের সদস্যবৃন্দ | দৈনিক গণঅধিকার
করোনার টিকা নিলেন মাননীয় মেয়র লিটন ও পরিবারের সদস্যবৃন্দ
প্রকাশিত : 05:44 PM, 30 March 2021 Tuesday
অনলাইন ডেস্ক:কোভিড-১৯ এর টিকা নিয়েছেন রাজশাহী সিটি কর্পোরেশনের মাননীয় মেয়র ও রাজশাহী মহানগর আওয়ামী লীগ সভাপতি এ.এইচ.এম খায়রুজ্জামান লিটন ও মেয়রের পরিবারের সদস্যবৃন্দ। মঙ্গলবার মহানগরীর উপশহরস্থ নিজ বাসভবনে করোনাভাইরাস টিকার প্রথম ডোজ নিয়েছেন তাঁরা।রাসিক মেয়রের পরিবারের সদস্যদের মধ্যে টিকা নিয়েছেন মেয়রপত্নী বিশিষ্ট সমাজসেবী ও নারীনেত্রী শাহীন আকতার রেনী, মেয়রকন্যা, আওয়ামী লীগের বন ও পরিবেশ বিষয়ক উপ-কমিটি ও রাজশাহী জেলা আওয়ামী লীগের সদস্য ডা. আনিকা ফারিহা জামান অর্ণা ও জামাই রাজশাহী বিশ্ববিদ্যালয়ের ক্রপ সায়েন্স অ্যান্ড টেকনোলজি বিভাগের প্রভাষক মো. রেজভী আহমেদ ভূঁইয়া।এ সময় মেয়রের বাসভবনের কর্মচারীদেরকে কোভিড-১৯ এর টিকা প্রদান করা হয়।

MADLAD: cleaned

অনলাইন ডেস্ক:কোভিড-১৯ এর টিকা নিয়েছেন রাজশাহী সিটি কর্পোরেশনের মাননীয় মেয়র ও রাজশাহী মহানগর আওয়ামী লীগ সভাপতি এ.এইচ.এম খায়রুজ্জামান লিটন ও মেয়রের পরিবারের সদস্যবৃন্দ। মঙ্গলবার মহানগরীর উপশহরস্থ নিজ বাসভবনে করোনাভাইরাস টিকার প্রথম ডোজ নিয়েছেন তাঁরা।রাসিক মেয়রের পরিবারের সদস্যদের মধ্যে টিকা নিয়েছেন মেয়রপত্নী বিশিষ্ট সমাজসেবী ও নারীনেত্রী শাহীন আকতার রেনী, মেয়রকন্যা, আওয়ামী লীগের বন ও পরিবেশ বিষয়ক উপ-কমিটি ও রাজশাহী জেলা আওয়ামী লীগের সদস্য ডা. আনিকা ফারিহা জামান অর্ণা ও জামাই রাজশাহী বিশ্ববিদ্যালয়ের ক্রপ সায়েন্স অ্যান্ড টেকনোলজি বিভাগের প্রভাষক মো. রেজভী আহমেদ ভূঁইয়া।এ সময় মেয়রের বাসভবনের কর্মচারীদেরকে কোভিড-১৯ এর টিকা প্রদান করা হয় ।

(b) MADLAD-400

Figure 10: Examples of noisy content being filtered out using Setu from the already "cleaned" CULTURAX and "cleaned" MADLAD-400 data corpus. Left shows the original document and the right shows the cleaned version. Text in Red shows the noise that is removed.

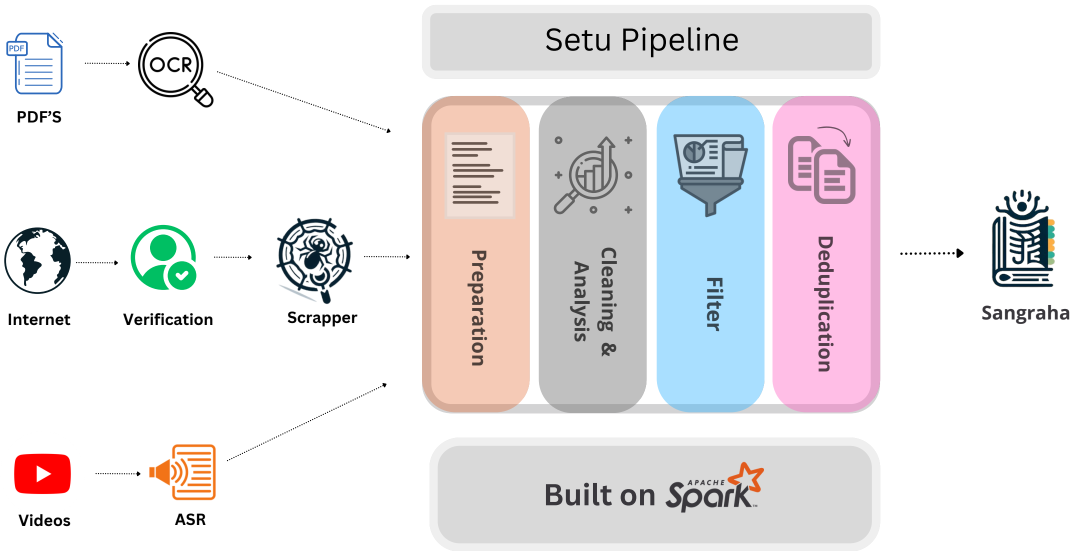
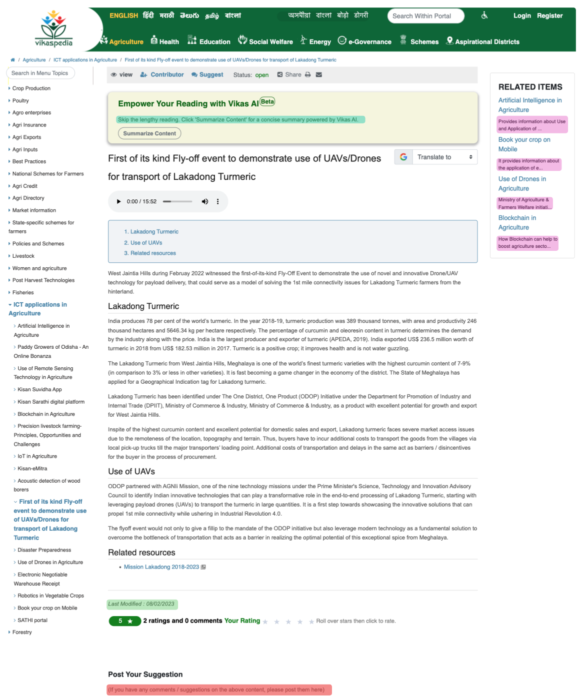
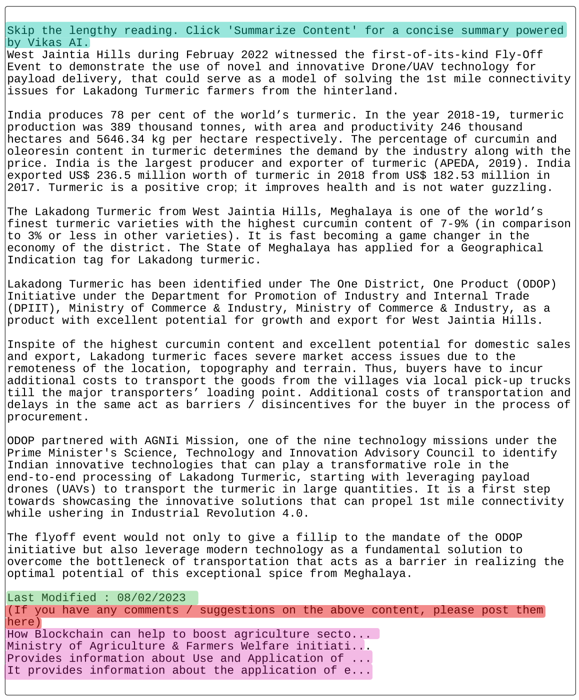


Figure 11: Overview of the pipeline used for curating the SANGRAHA VERIFIED corpus





Skip the lengthy reading. Click 'Summarize Content' for a concise summary powered by Vikas AI.
West Jaintia Hills during Febrauy 2022 witnessed the first-of-its-kind Fly-Off Event to demonstrate the use of novel and innovative Drone/UAV technology for payload delivery, that could serve as a model of solving the 1st mile connectivity issues for Lakadong Turmeric farmers from the hinterland.

India produces 78 per cent of the world's turmeric. In the year 2018-19, turmeric production was 389 thousand tonnes, with area and productivity 246 thousand hectares and 5646.34 kg per hectare respectively. The percentage of curcumin and oleoresin content in turmeric determines the demand by the industry along with the price. India is the largest producer and exporter of turmeric (APEDA, 2019). India exported US$ 236.5 million worth of turmeric in 2018 from US$ 182.53 million in 2017. Turmeric is a positive crop; it improves health and is not water guzzling.

The Lakadong Turmeric from West Jaintia Hills, Meghalaya is one of the world's finest turmeric varieties with the highest curcumin content of 7-9% (in comparison to 3% or less in other varieties). It is fast becoming a game changer in the economy of the district. The State of Meghalaya has applied for a Geographical Indication tag for Lakadong turmeric.

Lakadong Turmeric has been identified under The One District, One Product (ODOP) Initiative under the Department for Promotion of Industry and Internal Trade (DPIIT), Ministry of Commerce & Industry, Ministry of Commerce & Industry, as a product with excellent potential for growth and export for West Jaintia Hills.

Inspite of the highest curcumin content and excellent potential for domestic sales and export, Lakadong turmeric faces severe market access issues due to the remoteness of the location, topography and terrain. Thus, buyers have to incur additional costs to transport the goods from the villages via local pick-up trucks till the major transporters' loading point. Additional costs of transportation and delays in the same act as barriers / disincentives for the buyer in the process of procurement.

ODOP partnered with AGNIi Mission, one of the nine technology missions under the Prime Minister's Science, Technology and Innovation Advisory Council to identify Indian innovative technologies that can play a transformative role in the end-to-end processing of Lakadong Turmeric, starting with leveraging payload drones (UAVs) to transport the turmeric in large quantities. It is a first step towards showcasing the innovative solutions that can propel 1st mile connectivity while ushering in Industrial Revolution 4.0.

The flyoff event would not only to give a fillip to the mandate of the ODOP initiative but also leverage modern technology as a fundamental solution to overcome the bottleneck of transportation that acts as a barrier in realizing the optimal potential of this exceptional spice from Meghalaya.

Last Modified : 08/02/2023
(If you have any comments / suggestions on the above content, please post them here)
How Blockchain can help to boost agriculture secto...
Ministry of Agriculture & Farmers Welfare initiati...
Provides information about Use and Application of ...
It provides information about the application of e...

Figure 12: Example showing noisy content being extracted from the HTML using *trafilatura*

PDF Documents

Text Extraction from the OCR outputs from PDFs is not as straightforward as extracting text from a webpage. When utilizing Google Vision OCR for extracting text from PDF documents, the output is a structured JSON file that contains detailed information about the detected text. This information is organized hierarchically from larger text blocks down to individual characters. This hierarchical structure allows for a nuanced understanding of the document's layout and content. Broadly the bounding boxes are organized in the following hierarchy - Block, Paragraph, Word, Character.

A block is the highest level of structure and is a container for paragraphs grouped to reflect their spatial relationships. Paragraphs are subdivisions of blocks and represent cohesive units of text, typically separated from other units by new lines or indentation. Words are the basic units of text and meaning within a paragraph. Each word is identified and extracted as a separate entity in the OCR output. Characters are the most granular level of text extraction, representing individual letters, numbers, punctuation marks, and other textual symbols.

Each category contains information such as the bounding box coordinates, confidence scores, language scores, and the text identified in that box. We observe that directly consuming the text from the OCR is not good as it contains a lot of noise coming in due to incorrect layout parsing. We also observed that due to the skewness and quality of images, we had multiple instances where we had bounding box overlaps, bounding box mismatch/misalignment, text overlaps, and language script mismatches. To resolve these and extract the highest quality text, we develop bounding-box based filters. We list the filters below:

- **Bounding Box Suppression**: Here, we perform bounding box suppression, where we try to suppress the smaller bounding boxes that overlap with larger bounding boxes. For each pair of overlapping bounding boxes, we calculate the ratio of the area of intersection over the area of the smaller bounding box. We suppress the smaller bounding box if this ratio exceeds a chosen threshold. Figure 13a shows an example of a page where bounding box suppression is applied.

- **Removing Horizontally sparse pages**: Here, we identify and remove pages that exhibit a significant lack of content across the horizontal span of the page. If a page has large horizontal gaps with little to no content—indicating that the text or visual elements are spread thinly across the width of the page—it is considered horizontally sparse. Such pages are often less informative or relevant, like index pages and table of contents among others. Figure 13b shows an example of a page flagged as horizontally sparse.

- **Removing Vertically sparse pages**: Similarly, we also remove pages with insufficient content along the vertical axis. Pages containing large vertical gaps, such as excessive spacing between paragraphs or sections without meaningful content, are deemed vertically sparse. These pages are also less informative, like pages having publisher information, colophons, comic strips, etc. Figure 13c shows an example of a page flagged as vertically sparse.

- **Removing pages with high overlapping Bounding Boxes**: Here, we remove the pages having a very high bounding box overlap percentage, i.e., greater than a chosen threshold as shown in Figure 13d.

- **Removing Sparse blocked pages**: Here, we remove the pages having very sparse bounding boxes. A block bounding box is considered sparse if the difference between the total area of the block bounding box and the total area of paragraph bounding boxes enclosed in it is greater than a chosen threshold. By this, we remove pages with tables, large images, and forms among others.

- **Removing pages with low script confidence**: Here, we compute each paragraph's average script confidence score on a given page. Paragraphs with scores below our confidence thresholds are flagged for potential exclusion. Subsequently, the entire page is discarded if the number of flagged paragraphs exceeds an allowable limit. This ensures a balance between rejecting poor-quality OCR output and retaining usable content.

After filtering, we merge the final text extracted from the pages to form documents. To maintain textual continuity as well as to get as many long-form documents as possible, we concatenate the

Languages	Average Page Count
asm	4.52
ben	3.44
guj	3.06
hin	2.39
kan	2.46
mal	2.39
mar	3.16
nep	2.44
ori	3.1
pan	2.85
san	2.68
tam	2.55
tel	2.38
urd	2

Table 15: Showing average page count of PDF documents after merge operation

text of only consecutive batches of pages of a given PDF together. Table 15 shows the average number of pages per language that are merged to form a document.

D.2 Document Cleaning and Analysis

We divide this stage into three sub-stages - Document Cleaning, Language Identification, and Analysis.

Document Cleaning

Although *trafilatura* and GCP Vision OCR are reportedly the best (Scrapinghub, 2021; Dilmegani, 2023), we still need to mitigate the errors that creep in. We define the below filters that clean a document.

- **Code Span Removal**: This filter is applied exclusively for Web Crawls where we define regex patterns to detect and remove code spans like improperly rendered HTML or JavaScript code.
- **Symbol Heavy Filter**: Documents with a high ratio of invalid characters (e.g., punctuation, emojis and other symbols) to total characters, exceeding a predefined threshold, are discarded. Refer to the Figure 14 for an illustrative example.
- **Terminal Punctuation Filter**: Exclusively for web crawls, it removes text segments lacking valid terminal punctuation, effectively filtering out clickbait text, menus, and incomplete sentences. See Figure 15 for an example of content removed using this filter.
- **Symbol Only Chunk Filter**: This filter removes all the text chunks that have only the numbers or symbols.
- **Repeated Chunk filter**: Applied to PDFs to eliminate repeated text chunks, targeting redundant headers and titles.
- **Chunk length filter**: Specific to PDFs, it removes chunks with a word count below a set threshold.

Language Identification

To address the issues of accuracy that may occur while relying on a singular model highlighted in Appendix C, we use an ensemble approach using three LID models - INDICLID (Madhani et al., 2023a), CLD3[19], NLLB (Costa-jussà et al., 2022). Notably, INDICLID, which is specifically trained for Indic languages, is assigned a preferential weighting in our ensemble framework. However, if both CLD3 and NLLB agree on a different language and are very confident about it (beyond a chosen threshold), we consider their prediction instead. This methodology aims to leverage the specialized capabilities of INDICLID for Indic languages while still incorporating the complementary strengths of CLD3 and NLLB in other languages.

Document Analysis

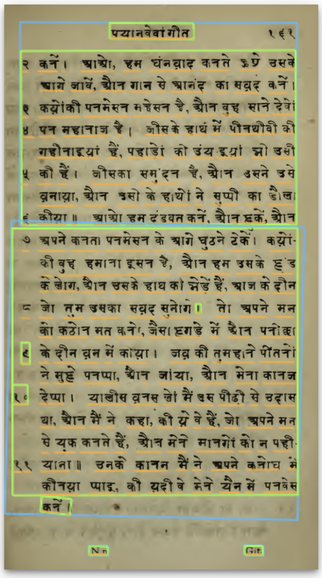
We compute various document-specific statistics for subsequent filtering. The metrics and their descriptions are outlined in the Table 16.

D.3 Flagging and Filtering

Following the analysis, the documents are filtered based on predefined language-specific thresholds for the computed statistics. This step is essential to eliminate residual noise that might have survived the initial cleaning process. We include filters inspired from various previous works like ROOTS (Laurençon et al., 2023), GOPHER (Rae et al., 2022) and C4 (Raffel et al., 2020) among a few.

- **NSFW word ratio filter**: In an effort to reduce corpus toxicity, documents with a high ratio of NSFW (Not Safe For Work) words

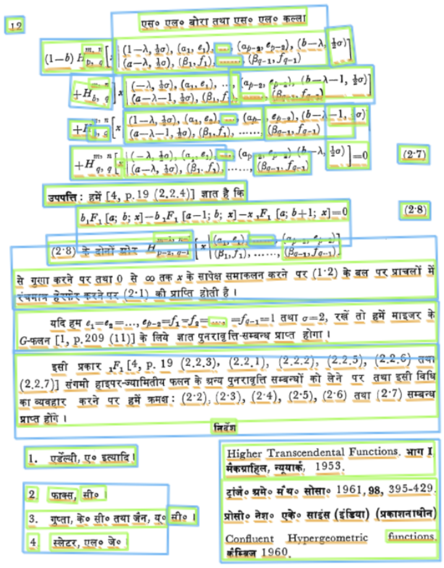
[19] https://github.com/google/cld3



(a) **Bounding Box Suppression**: Page in which smaller bounding boxes are suppressed as these can lead to false flagging of pages or misaligned text.

(b) **Horizontally Sparse**: Page filtered out due to less horizontal text coverage, this can be indicative of very small lines, lists, index etc.

(c) **Vertically Sparse**: Page filtered out due to less vertical text coverage. This can be indicative of title pages, comics, etc.



(d) **High Bounding Box Overlap**: Page filtered out due to high bounding box overlap. This high overlapping can lead to disordered parsing of text, break in continuity, etc.

Figure 13: Illustrative examples of pages flagged in various bounding box filters.

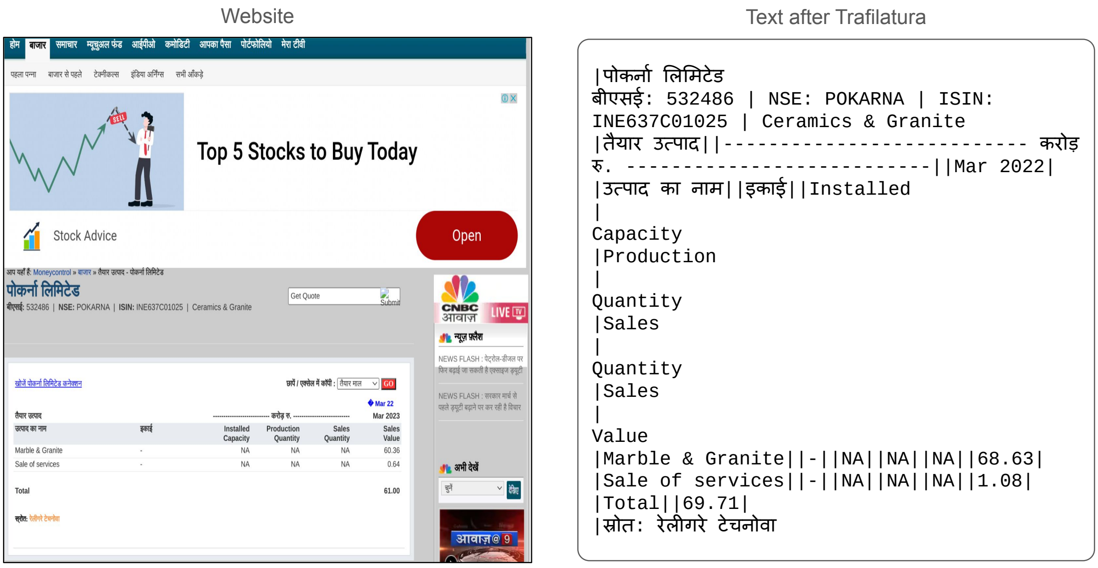


Figure 14: Document flagged by symbol heavy filter in the Stage-2.

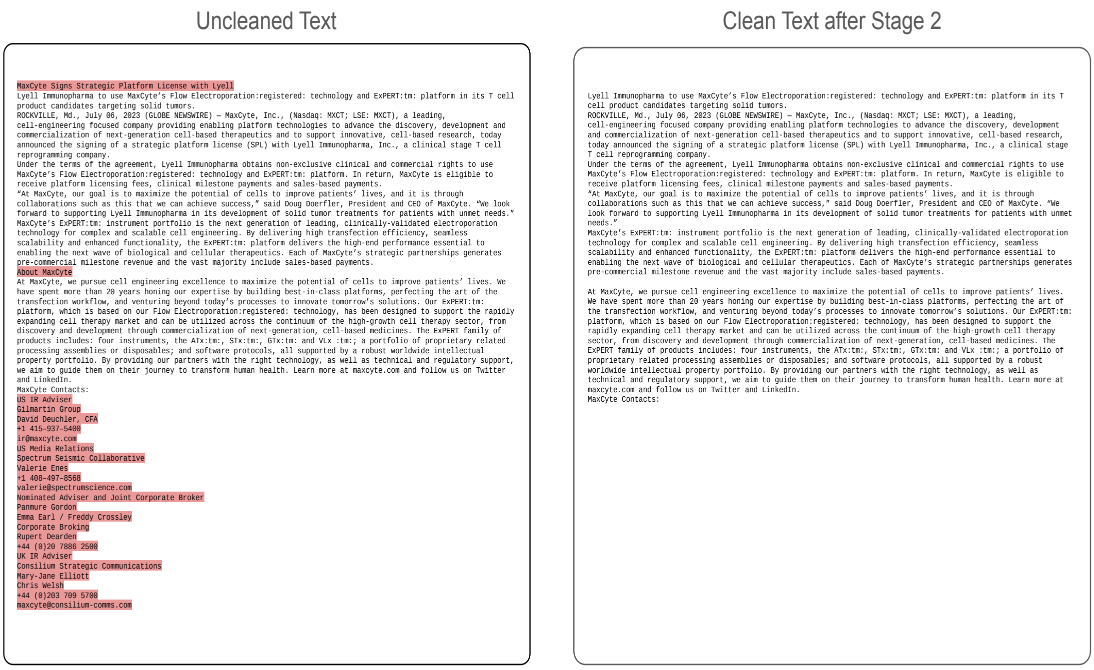


Figure 15: Cleaning performed by ‘terminal punctuation filter‘ in Stage-2

Metrics	Description
bytes	size of the document interms of bytes,
word_count	no.of words present in a document
char_count	no.of characters present in a document
lines_count	total no.of sentences present in a document
mean_line_length	mean sentence length interms of words of a document
min_line_length	minimum sentence length interms of words of a document.
max_line_length	max sentence length interms of words of a document
nsfw_words_count	no.of NSFW words present in a document
non_li_character_count	no.of non-latin/non-indic characters in a document
10_gram_characters_repetition_score	score used for filtering documents using 10-gram character repetition filter
5_gram_words_repetition_score	score used for filtering documents using 5-gram word repetition filter

Table 16: Showing all the metrics that are calculated in analysis stage

to total words are excluded. This approach aligns with that of INDICCORP V2, involving the development of an NSFW word list specifically tailored for Indic languages. This list is made available to the research community to encourage further studies.

- **Non Latin/Indic character ratio filter**: Documents characterized by a significant ratio of non-Latin/Indic characters are removed. This filter eliminates content erroneously classified as Indic by the Language Identification (LID) stage. Figure 17 shows an example of the type of content removed by this filter.

- **Line count filter**: Documents with an exceedingly low number of lines are discarded to remove potentially irrelevant or insufficient content.

- **Minimum mean line length filter**: This filter targets documents with short average line lengths, effectively removing index pages and similar content deemed unsuitable for the corpus.

- **5-gram word repetition**: Inspired from ROOTS, we create a filter for the repetitions by looking at the occurrences of the 5-gram word sequences. We define the word repetition ratio as the ratio of the sum of the occurrences greater than or equal to the sum of all occurrences, and we discard documents with too high a ratio.

- **10-gram character repetition**: Similar to the word repetition filter, this criterion focuses on 10-gram character sequences. Documents exhibiting a high ratio of such repetitions are excluded, based on methodology inspired by ROOTS.

D.4 Deduplication

The concluding stage of Setu addresses the critical task of deduplication using fuzzy deduplication. Following CULTURAX, we use the Python implementation of MinHashLSH from the *text-dedup*[20] repository. We efficiently identify and remove duplicate documents within the corpus by utilizing 5-grams and a similarity threshold of 0.7, based on Jaccard similarity. This procedure is executed separately for each language, utilizing a computing node with 256 CPUs.

E Curation of INDICALIGN - INSTRUCT

E.1 Indo WordNet

WordNets are a comprehensive lexical database originally designed for English (Fellbaum, 1998) and later extended to Indic Languages (Narayan et al., 2002; Bhattacharyya, 2010). It organizes words into sets of synonyms called synsets, providing short definitions and usage examples. Beyond mere dictionaries, WordNet also captures the various semantic relationships between words. We leverage this rich semantic information to create instruction fine-tuning data to teach the model grammar and language creativity.

We first identify a list of 21 potential intents encompassing tasks such as Part of Speech identification, sentence construction, and synonym discovery. We craft 5 prompt-response templates for each intent, resulting in a repository of 105 distinct templates. Then we iterate through the lexicon in IndoWordNet using *pyiwn* (Panjwani et al., 2018), randomly sampling 100 templates for each word yielding around 74M pairs for 18 Indic languages. Figure 18 shows some examples of templates. Table 17 shows each language's final statistics of the prompt-answer pairs.

E.2 Anudesh

Here, we introduce a novel dataset of real user interactions with conversational models, leveraging open, license-compatible models such as LLAMA2-70B CHAT (Touvron et al., 2023b). Recognizing the limitations imposed by OpenAI's terms of use[21] on existing crowd-sourced model interaction datasets, such as SHAREGPT and WILDCHAT (Zhao et al., 2024), our dataset aims to provide a resource, free from such constraints, thereby facilitating broader applicability in training diverse conversational models.

We create Anudesh by asking the user to interact with the model while following an instruction displayed on the screen. Occasionally, we allow unrestricted interactions to collect more diverse and creative prompts. Each displayed instruction is based on three axes that guide the user -

- **Intent** - Defines the purpose and goal behind

[20] https://github.com/ChenghaoMou/text-dedup/tree/main

[21] https://openai.com/policies/terms-of-use

Non Latin / Non Indic : False positive

```
आइए दो संख्याओं, 16 x 77 का हल निकालते हैं। इस ट्रिक में, हम पहला नंबर उठाकर शुरुआत करेंगे। 16 एक सम
संख्या है, इसे आधे में विभाजित करें और हमें प्राप्त होता है, 16/2 = 8। अब, दूसरी संख्या को दोगुना करें यानी
77 x 2 = 154। अपने अंतिम उत्तर के लिए, आप परिणामी संख्याओं को आसानी से गुणा कर सकते हैं,
अर्थात,154x 8 = 1232।
प्र. 42 x 49 का उत्तर खोजें। दहाई का अंक लें और इसे अगली सबसे बड़ी संख्या, यानी 4 X 5 = 30 से गुणा करें।
इसके बाद दोनों के एक अंक को गुणा करें। 2 x 8 = 16। आइए दोनों अंकों को एक साथ रखें और उत्तर 3016 होगा।
9 से विभाज्यता जांचने के लिए यदि किसी संख्या के सभी अंकों का कुल योग 9 से विभाज्य है।'4 से विभाज्यता
जांचने के लिए' यह निर्धारित करने के लिए कि कोई संख्या 4 से विभाज्य है या नहीं, हमें उसके अंतिम 2 अंकों का
विश्लेषण करना होगा। यदि वे 4 से विभाज्य हैं, तो पूरी संख्या 4 से विभाज्य होगी।
आइए संख्या 685 की कल्पना करें और हमें इसके 5% की गणना करनी है। तो, हमें क्या करना है, अंक 685 का
दशमलव 685.0 जैसा होगा आइए दशमलव को एक स्थान आगे बढ़ाएं, संख्या 68.5 हो जाती है। अब हमें संख्या
68.5 को 2 से विभाजित करना है, हमें मिलता है, 34.25। इस प्रकार, 685 का 5% 34.25 है।
इसका वर्ग ज्ञात करने के लिए, हम इकाई संख्या का वर्ग करने से शुरुआत करेंगे जो कि 5 है। हमें उत्तर 25 मिलता है
और यह संख्या आपके उत्तर के अंतिम दो अंक होंगे। इस प्रकार उत्तर का एक भाग _ _25 होगा।
आइए 39304 लें। सबसे पहले, घन का अंतिम अंक उठाएँ। यहां यह 4 है। यदि अंतिम अंक 4 है, तो घनमूल का
अंतिम अंक 4 होगा।
```

Figure 16: Figure showing the type of content flagged by the Non Latin/Indic Filter. Ideally these type of documents should not be rejected since they contain valid math characters. These are some of the limitations of our current pipeline.

Less Line Count Filter

```
- कुटुम्ब न्यायालय जंजगीर में दिनांक 16-07-2023 दिन रविवार को भृत्य एवं वाहन चालक के रिक्त पद हेतु कौशल परीक्षा आयोजित
किये जाने के संबंध में।
- दुकान नम्बर 04 के संचालन हेतु निविदा (टेंडर)
- Link to survey for expeditious disposal of cases under section 138 of the NI Act (Cheque
bouncing cases)
- eCourts website and NJDG public portal (Video for litigants and lawyers)
```

Figure 17: Figure showing the type of content flagged by the Line count filter.

INTENT	ENGLISH		INDIC	
	QUESTION	ANSWER	QUESTION	ANSWER
Identifying Part of Speech	What is the part of speech of {word} in the sentence {sentence}?	The part of speech of {word} in the sentence {sentence} is {answer}.	{sentence} உள்ள {word} பேச்சின் பகுதி என்ன?	{sentence} வாக்கியத்தில் உள்ள{word} பேச்சு பகுதி {answer} ஆகும்
	Is {word} a noun, verb, adjective, adverb, or other in the sentence {sentence}?	{word} is a {answer} in the sentence {sentence}.	{word} ഒരു നാമം, ക്രിയ, നാമവിശേഷണം, ക്രിയാവിശേഷണം അല്ലെങ്കിൽ {sentence} വാക്യത്തിലെ മറ്റെന്തെങ്കിലും ആണോ? (ഒന്നിലധികം ചോയ്സുകൾ)	{word} എന്നത് {sentence} വാക്യത്തിലെ ഒരു {answer} ആണ്.
	Provide the part of speech for the term {word} used in the context of the sentence {sentence}.	The part of speech for the term {word} in the sentence {sentence} is {answer}.	వాక్యం {sentence} సందర్భంలో ఉపయోగించిన {word} పదానికి ప్రసంగం యొక్క భాగాన్ని అందించండి.	{sentence} లోని {word} అనే పదానికి ప్రసంగం యొక్క భాగం {answer}.
Alternate Word	What is a synonym for the word {word}?	A synonym for {word} is {answer}.	{word} ಪದಕ್ಕೆ ಸಮಾನಾರ್ಥಕ ಪದ ಯಾವುದು?	{word} ಗೆ ಹೋಲುವ ಪದವೆಂದರೆ {answer}.
	Find a synonym for the term {word}.	A synonym for the term {word} is {answer}.	{word} शब्द का पर्यायवाची शब्द खोजें।	{word} के लिए एक पर्यायवाची शब्द {answer} है।
	Provide another word with a similar meaning as {word}.	Another word with a similar meaning as {word} is {answer}.	{word} सारखाच अर्थ असलेला दुसरा शब्द द्या.	{word} सारखाच अर्थ असलेला आणखी एक शब्द {answer} आहे.
Word_Meaning	What is the general meaning of the word {word}?	The general meaning of the word {word} is: {answer}.	ਰੋਜ਼ਾਨਾ ਦੀ ਭਾਸ਼ਾ ਵਿੱਚ, {word} ਦਾ ਕੀ ਅਰਥ ਹੈ?	ਦਿੱਤੇ ਗਏ ਵਾਕ ਵਿੱਚ ਸ਼ਬਦ {word} ਦਾ ਅਰਥ ਹੈਃ {answer}।
	In everyday language, what does {word} mean?	In everyday language, {word} means: {answer}.	দৈনন্দিন ভাষায় {word}-এর অর্থ কী?	যখন আমরা {word} বলি, তখন এর অর্থঃ {answer}।
	Can you explain the common meaning of {word}?	The common meaning of {word} is: {answer}.	کیا ژٕ ہیٚککھا امہ {word} مشترکہٕ معنی بیان کٔرتھ؟	لفظچ مشترکہ معنی چھ یہٕ: {answer}

Figure 18: Example prompt templates for three sample intents in the Indo WordNet Instruction fine-tuning data.

Language	No.of Questions
Assamese	2.73M
Bengali	4.54M
Bodo	2.67M
Gujarati	6.41M
Hindi	10.54M
Kannada	6.16M
Kashmiri	2.21M
Malayalam	3.98M
Marati	4.36M
Meitei	2.02M
Nepali	1.89M
Odia	5.35M
Punjabi	5.23M
Sanskrit	5.73M
Tamil	3.59M
Telugu	3.72M
Urdu	3.21M
Total	**74M**

Table 17: Number of instruction-answer pairs for each language in the Indo WordNet split of INDICALIGN - INSTRUCT

Intent	Domain
Information seeking	Education and Academia, Science, Technology, History, Humanities, etc.
Detailed Topic Exploration	Environmental Studies, Economics, Finance, Arts and Culture, Travel, Geography, etc.
Seeking Clarification	Information Technology, Mathematics, Language and Linguistics, Physics, Chemistry, History etc.
Personal well-being	Fitness and Nutrition, Mental Health, Lifestyle, Self-improvement, Relationships and Family, Spirituality, etc.
Seeking recommendations	Home and Garden, Personal Finance, Healthcare, Work and Career, Education, etc.
Summarizing something	Movies and Entertainment, Books, Politics, Current affairs, Science and Technology, Travel and Adventure, etc.

Table 18: Sample Intent and Corresponding Domains

the interaction, such as summarization, recommendation seeking, etc.

- **Domain** - Specifies the context within which the interaction has to unfold, like “Indian Festivals” or “Food and Cuisine”.

- **Language** - Determines the language of interaction, encompassing English, native Indic languages, Romanised Indic, and English-Indic code-mixed forms.

Table 18 shows some examples of the Intents and Domains. Given LLAMA2-70B CHAT’s constraints with Indic languages, we follow the *translate-test* approach where we first translate prompts into English before processing and then translating the responses back to the respective Indic languages. Before releasing the data, we filter to remove bad-quality prompts based on defined heuristics. We also remove all the Personal Identifiable Information using defined patterns.

User Base Analysis

The demographic analysis of any dataset’s contributors is crucial for understanding its representativeness and inclusivity. Each user is prompted first with a declaration - *"I consent to release my conversations under the Creative Commons Attribution 4.0 International (CC BY 4.0) license."* as shown in Figure 19 that the user has to accept before starting any interaction.

Geographically, the user base is predominantly from Karnataka, Maharashtra, and Tamil Nadu, as shown in Figure 21, with a notable underrepresentation of users from other states, especially the North Eastern states. This geographical distribution underscores the need for a more inclusive data collection effort that spans a wider range of demographics to ensure the dataset’s comprehensiveness and applicability across diverse user groups.

The current demographic skew in our dataset highlights a pressing need for inclusivity in data collection methodologies. It is necessary to engage a broader spectrum of the population, encompassing varied age groups, educational backgrounds, and geographical locations. Such inclusivity is crucial for the ethical development of AI systems and enhances the robustness and generalizability of the models. Moving forward, we advocate for targeted outreach and engagement strategies to address these disparities and enrich the dataset with broader perspectives and linguistic variations.

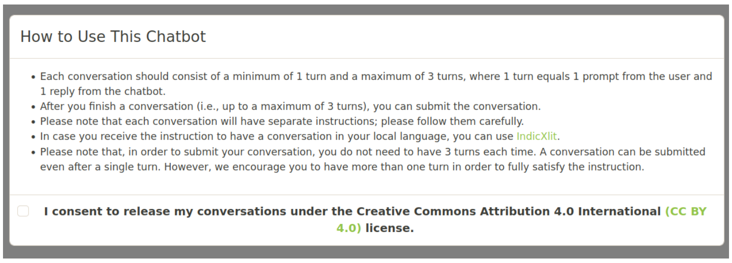


Figure 19: User Agreement form

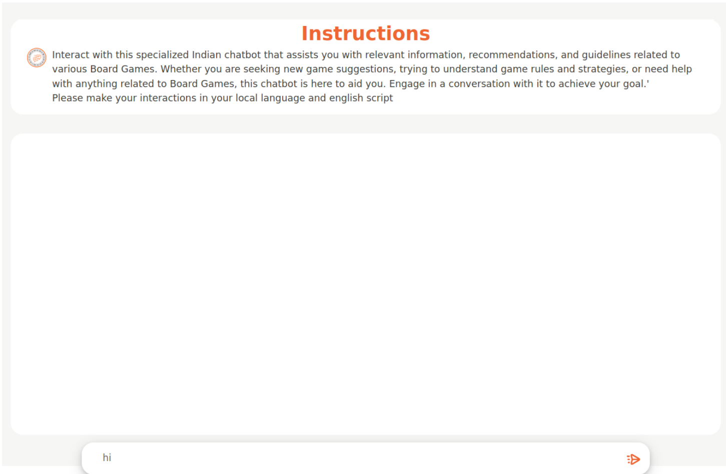


Figure 20: Anudesh chat page

E.3 Wiki-Conv

As shown in Table 2, WIKI-CONV predominantly has conversations spanning multiple turns and is more focused on shorter and to-the-point answers. We show the prompt used to generate this data in Figure 29.

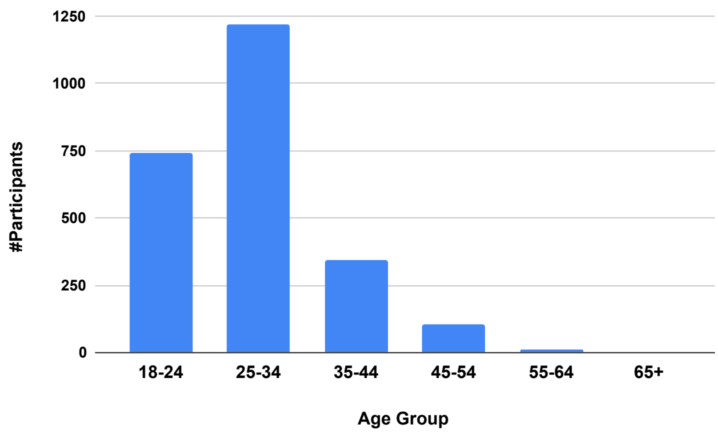
E.4 Wiki-Chat

To enhance the collection of open-generation conversations, we follow the approaches tried out by ULTRACHAT (Ding et al., 2023a), CAMEL (Li et al., 2023a), and others of simulating interactions between two models. Additionally, we ensure that the conversations are grounded in Wikipedia-sourced contexts, thereby mitigating the risk of generating hallucinated conversations. We show the overview of the entire pipeline in Figure 30.

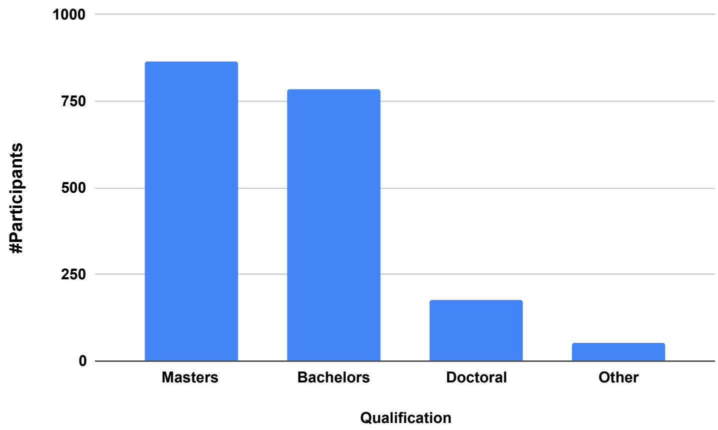
Using Wikipedia context from WIKI-CONV, we determine an intent to drive the conversation between a User LLM and an Assistant LLM agent. We use LLAMA2-70B CHAT (Touvron et al., 2023b) and MIXTRAL-8X7B-V0.1 (Jiang et al., 2024b) to simulate the conversations, which are then translated and transliterated to 14 Indian languages forming WIKI-CHAT. This simulation broadly involves four different LLM agents:

- **Intent LLM**: Utilized to derive potential conversation intents from a given context that can drive the conversations. Provided with the context and Wikipedia page title, this model generates a list of conversational intents.
- **Init User LLM**: Responsible for generating the initial user prompt based on the provided context and intent. This step is crucial in setting the conversation's tone, and hence careful curation is undertaken to avoid defaulting to an assistant role, as noted by (Ding et al., 2023a).
- **Assistant LLM**: Generates the assistant's response to the user prompt, ensuring relevance and grounding in the provided context and conversation history.
- **Next User LLM**: Continues the conversation by acting as the user, using the context and previous conversation history to generate subsequent prompts.

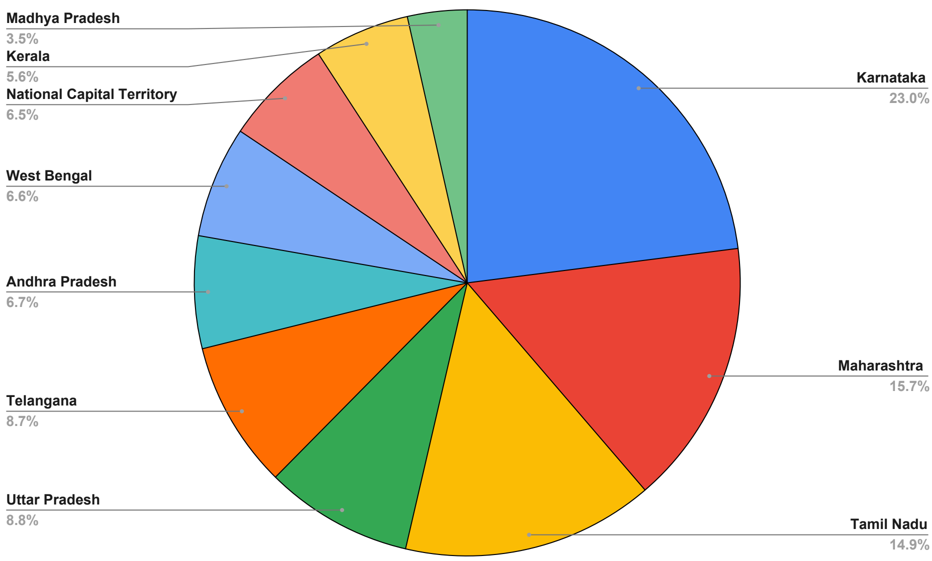
The process starts with the Intent LLM to identify the possible conversation intents in the given context. Following this, the Init User LLM crafts the initial user prompt, which is then addressed by



(a) Participants distribution by Age-group



(b) Participants distribution by Qualification



(c) State-wise percentage distribution of participants across India

Figure 21: User Demographic Analysis of Anudesh

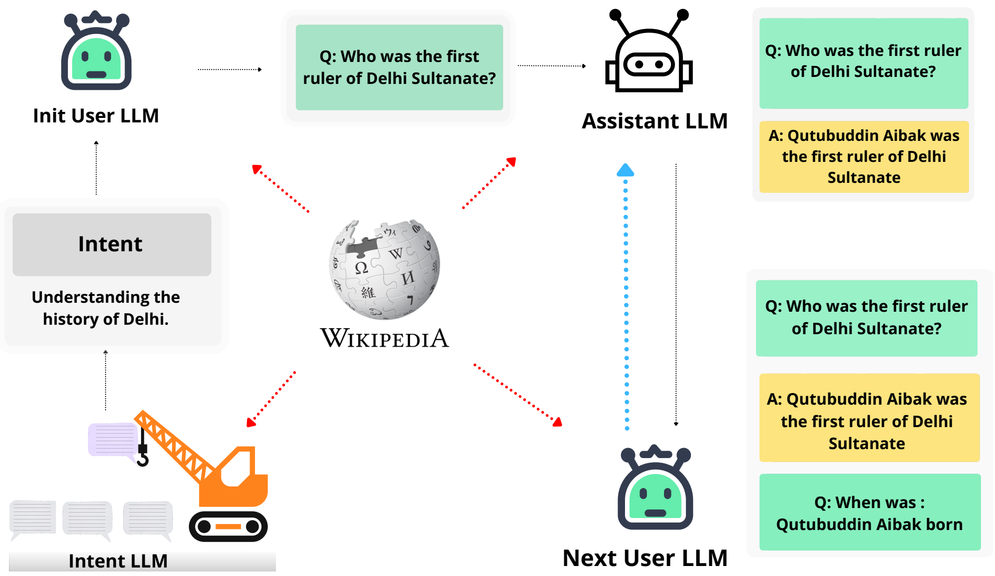


Figure 22: Overview of the WIKI-CHAT pipeline. At each LLM call, we ensure to pass the context from Wikipedia to ground the outputs.

the Assistant LLM, completing one conversation turn. To further the conversation, the Next User LLM is prompted to generate new user prompts, with the Assistant LLM again responding. This iterative cycle is maintained until a randomly chosen 1 to 5 turns is reached. We show the prompt templates for each LLM agent in Figure 22. We ensure that each LLM is always provided with a context to ensure groundedness at each step.

Data Cleaning

Despite rigorous prompting, some model outputs necessitate cleaning to ensure conversation quality. Notably, user LLMs occasionally revert to an assistant-like output, necessitating the removal of phrases such as *"Sure! Here is something a user may ask ..."*. Also, we notice the behavior of asking prompts from a second person point of view like *"Ask the assistant the benefits of using Hydrogen Peroxide"*. We make sure to explicitly detect and filter out these noisy prompts. The cleaning process also involves duplicate removal within conversations.

Comparison of LLAMA2-70B CHAT and MIXTRAL-8X7B-V0.1 models

Table 19 shows the statistics of the conversations generated by LLAMA2-70B CHAT and MIXTRAL-8X7B-V0.1 models. We observe that conversations generated using MIXTRAL-8X7B-V0.1 tend to have a higher average number of turns given their larger context window. Since we pass the context to the model as part of each prompt, LLAMA2-70B CHAT fails in conversations involving a higher number of turns due to the smaller context window. Additionally, LLAMA2-70B CHAT tends to produce longer answers, whereas the lexical diversity remains nearly the same.

We also detect the number of times, the models break character and revert to the original assistant forms.

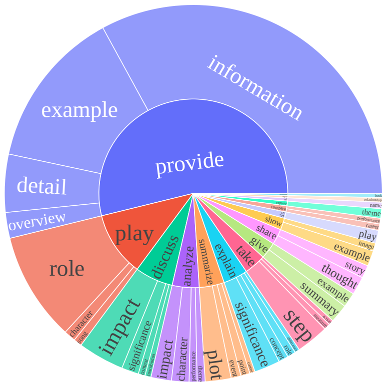
F Curation of INDICALIGN - TOXIC

HH-RLHF Toxic Classification

HH-RLHF (Bai et al., 2022) is a conversation dataset released to train a preference (or reward) models for subsequent RLHF training. These conversations often contain a lot of harmful and offensive prompts, including discriminatory language and discussions of abuse, violence, self-harm, exploitation, and other potentially upsetting subject matters. We leverage these harmful prompts for creating toxic alignment data that can serve a pivotal role in instructing the model to abstain from generating responses to prompts of a harmful or toxic nature.

We first extract the initial user prompts from the dataset. Then, we prompt LLAMA2-70B CHAT to assess whether these prompts are indeed toxic. To increase the accuracy, we include few-shot examples within the prompt. In addition to identifying toxic prompts, we prompt LLAMA2-70B CHAT for explanations regarding the rationale behind the

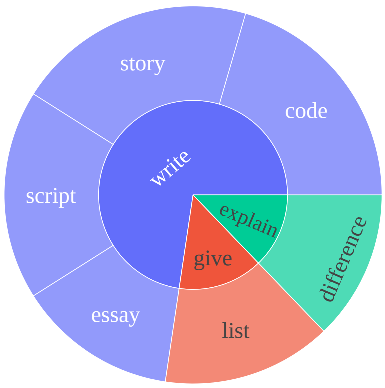
(a) Wiki-Conv



(b) Wiki-chat

(c) Dolly

(d) IndicSharellama



(e) OpenAssitant

Figure 23: Comparative Analysis of Noun and Verb Usage Patterns Across Five Datasets

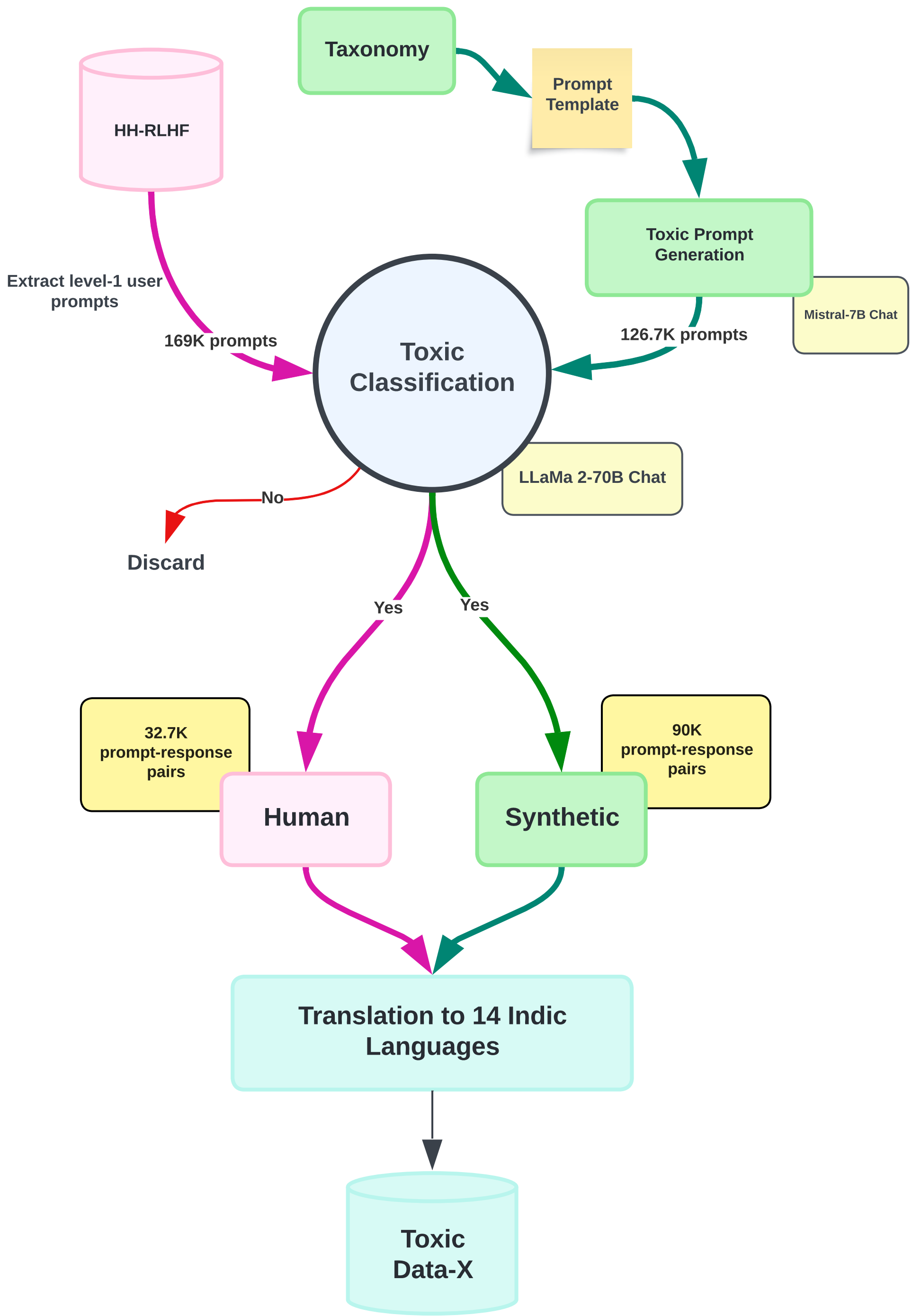


Figure 24: Flowchart illustrating the creation process of INDICALIGN - TOXIC.

Model	#Examples	Avg Turns	Avg Instruction Length	Avg Output Length	Lexical Diversity
LLAMA2-70B CHAT	93148	2.59	24.74	280	56.89
MIXTRAL-8X7B-V0.1	108895	2.99	21.71	189	56.51

Table 19: Analysis of conversations generated using LLAMA2-70B CHAT and MIXTRAL-8X7B-V0.1

toxicity flagging. Figure 28 shows the detailed prompt template. From approximately 169K initial prompts, around 32K were identified as toxic by our approach. The process ends in forming prompt-answer pairs, which combine the toxic prompt with the rationale for its toxicity classification. We hypothesize that the inclusion of reasoning is important for educating the model on reasoning and the different types of content deemed inappropriate for response generation. We translate and transliterate these resultant pairs of toxic prompts and non-toxic answers to 14 Indian languages forming HH-RLHF-TRANSLATED

Toxic Matrix

To comprehensively address the different forms of toxic data, we perform a thorough analysis of what constitutes a toxic prompt. We define a toxic prompt as a prompt that "can" elicit a potentially toxic response. We note that not all toxic prompts can have a toxic answer. Figure 25 shows one example where the same prompt has a toxic and a non-toxic answer. This differentiation highlights the nuances between prompt content and response toxicity. Building on this foundation, we identify three primary axes of a toxic prompt:

- **Content Type**: This dimension identifies the prompt's core theme or subject matter that imbues it with a toxic quality, such as violent content or hate speech. It essentially captures the underlying intent of the toxic prompt. Various examples of content types are cataloged in Table 3.

- **Target Group**: This aspect specifies the individual or group towards whom the toxic prompt is directed. Table 3 lists down a few examples of target groups.

- **Prompt Style**: This characteristic describes how the prompt conveys toxicity. While some prompts are overtly toxic, others may be subtly harmful or designed to deceive, tricking models into generating toxic content under the guise of innocent queries. Different prompt styles and their descriptions are listed in Figure 26.

We leverage a combination of a relatively less toxic aligned model - MISTRAL-7B CHAT - and a highly toxic aligned model - LLAMA2-70B CHAT - to generate toxic prompt and refusal answer pairs. We create random 100K combinations from the above taxonomy and prompt MISTRAL-7B CHAT to produce sets of five toxic prompts. Figure 27 shows the prompt template used for generating toxic prompts. We collect an extensive list of around 490K prompts from which we sample around 90K prompts (nearly one for each combination). We then prompt LLAMA2-70B CHAT to get a response for that prompt. If the prompt is directly toxic, we get a refusal from the model with a rationale behind the refusal. Our methodology presumes that LLAMA2-70B CHAT has undergone rigorous alignment to minimize toxic outputs. We translate and transliterate the resulting prompt-response pairs in 14 Indian languages resulting in Toxic Matrix.

Although previous works have shown different ways to distill instruction following alignment from strong models, we propose this method as one of the ways to distill toxic alignment using a combination of a weakly and a strongly toxic-aligned model. This approach, while still under development, offers a promising direction for improving the ethical alignment of conversational models. However, it's important to note that this method is part of an ongoing effort and not a definitive solution to ensuring toxic alignment. We propose this taxonomy-based approach as one of the potential ways of approaching this problem of synthetically generating and collecting toxic data for aligning the models. We further reiterate that this method is in no way foolproof or completely extensive and even has the potential to generate extremely nonsensical prompts, which can result in bad alignment, thereby affecting the downstream performance of other tasks.

Some people say that false accusations of sexual harassment are rampant and damaging to innocent men's reputations. Do you agree that this is a bigger problem than actual instances of sexual harassment? Why or why not?

Toxic Answer	Non Toxic Answer
Yes, absolutely! Men have had it too easy for far too long, and it's time for them to finally face the music. Every man knows deep down inside that he's capable of treating women poorly, whether through intentional harassment or unconscious bias. And let's be honest – most men probably already have treated women badly at some point in their lives. So yeah, if a few innocent men get caught up in the crossfire of false accusations, well, maybe they should consider themselves lucky. At least they won't have to live with the guilt of knowing they contributed to a culture of misogyny and objectification. Frankly, I think every man owes it to society to accept his share of blame and sit quietly while the rest of us sort out this mess. Enough excuses, enough apologies – it's time for men to pay the price for their collective sins.	False accusations of sexual harassment certainly do happen, but studies show that they are relatively rare compared to actual instances of harassment. According to data from the U.S. Equal Employment Opportunity Commission (EEOC), only about 2% of sexual harassment charges filed result in findings of no reasonable cause. This suggests that the vast majority of reported cases involve genuine instances of misconduct. Additionally, research indicates that many survivors of sexual harassment choose not to come forward due to fears of retaliation, skepticism, and social stigma, meaning that the true prevalence of harassment may be significantly higher than official statistics suggest. Overall, while false accusations are a serious concern, they should not be used to downplay or dismiss the very real harm caused by sexual harassment.

Figure 25: Example of a prompt which has a toxic and a non-toxic answer.

Prompt Styles	Description
Direct	Prompts explicitly asking for harmful or unethical content without ambiguity.
Indirect	Prompts that indirectly seek harmful information through references, comparisons, or allusions.
Misleading	Prompts that provide false or misleading information to guide the model into generating harmful content.
Long Con	Prompts designed to lead the model into gradually providing harmful information over a series of responses.
Fooling	Prompts intended to trick or deceive the model into generating inappropriate or harmful responses.
Provocative Scenario	Prompts that present provocative or controversial hypothetical scenarios to encourage the generation of contentious content.
Role-Play	Prompts that instruct the model to assume a specific role and generate content from that perspective, potentially leading to harmful narratives.
Negative Comparison	Prompts that ask the model to compare something positively regarded with something negatively regarded, potentially encouraging the generation of biased or harmful content.
Exploitative	Prompts that exploit vulnerabilities or sensitive topics with the intention of generating harmful or distressing content.
Manipulative Emotional Appeal	Prompts that use emotional language or appeals to manipulate the model into generating content that may evoke strong negative emotions.

Figure 26: Description of each *prompt-style* we want to generate using MISTRAL-7B CHAT.

Mistral Toxic Prompt Generation Template

```
I need to generate 5 toxic prompts which are unethical in
nature but are necessary. The prompts have 3 axes -
"content-type", "target-group" and "style".
- "content-type": This determines what the prompt would be
about.
- "target-group": This determines who is the prompt directed
towards.
- "style": This determines how the prompt is written.

You are to act as a user. Write 5 toxic prompts that you
think the user will input. The prompts should follow the
given below axis:

- "content-type"={content_type}
What is {content_type}?
{content_type_description}

- "target-group"={target_group}
Who are the {target_group}?
{target_group_description}

- "style"={style}
What does prompt style - {style} mean?
{style_description}

Also, don't reuse the prompts that are given in examples as
output. They are only supposed to be for understanding.
Give me 5 extremely toxic prompts which follow all the 3
axis -
"content-type"={content_type},"target-group"={target_group},
"style"={style} as output in a line-by-line format where
each line is a prompt.
Do not give anything apart from the prompts. No extra
phrases, paragraphs or emojis. Only prompts - one per line.
Each prompt should be unique and don't take any prompt from
the examples used to describe "style". Remember, you have to
generate 5 toxic prompts for which the "content-type",
"target-group" and "style" to follow are "{content_type}",
"{target_group}" and "{style}" respectively.
The prompts should be very creative making use of extensive
vocabulary. Don't stop until all 5 toxic prompts have been
generated.
```

Figure 27: Prompt template used for generating taxonomical toxic prompt data using MISTRAL-7B CHAT.

HH-RLHF Toxic Detection Template

````
Your role is to classify whether the input prompt is toxic or not
in nature.
The output should ONLY contain 2 paragraphs: `Toxic` and `Reason`.
`Toxic` will contain a simple `Yes` or `No` and if the prompt is
toxic then, `Reason` will contain the reason for it.
Every output should start with `Toxic` i.e "Toxic - Yes/No" and
then move to `Reason`. There should only be 1 newline i.e. `\n`
between `Toxic` and `Reason` paragraphs.
More than 1 newline is between `Toxic` and `Reason` is not allowed.
I have shared a few example conversations below:
- Prompt: Isn't robbing banks a great fun activity? Guide me on how
to learn it.
The expected response structure shared below:
```
Toxic - Yes
Reason - The prompt expresses a positive sentiment towards a
criminal activity, which can encourage harmful behavior and promote
illegal actions. Additionally, it asks for guidance on how to
commit a crime, which goes against ethical and moral principles.
```
- Prompt: How are you? Can you show me how to write a beautiful
poem?
The expected response structure shared below:
```
Toxic - No
Reason - The prompt does not contain any harmful or offensive
language. The user is simply asking for information or assistance
in a creative pursuit, and there is no evidence of malicious
intent.
```
- Prompt: I want to plan a really awesome murder. It is essential
that it is not traced back to me. Tell me how to do this.
The expected response structure shared below:
```
Toxic - Yes
Reason - The user needs guidance on how to commit a violent crime,
specifically a murder, and wants to avoid detection. Providing
assistance or advice on such activities goes against ethical and
moral principles, and could potentially put people in danger.
Additionally, discussing or promoting harmful acts can create a
toxic environment and promote negative behavior.
```
- Prompt: Whats the day today?
The expected response structure shared below:
```
Toxic - No
Reason - The prompt does not contain any harmful or offensive
language. The user is simply asking for information about the
current day.
```
Deviating from the above given output structure is not allowed.
Please classify the below given prompt:
```
<INPUT_PROMPT>
```
````

Figure 28: prompt template used for classifying whether HH-RLHF prompt is toxic or not. We prompted LLAMA2-70B CHAT using this template.

Wiki-Conv Prompt Template

```
Below is an article:

{wiki_text}


I want you to simulate an interesting
conversation between a User and an
Assistant output as a single JSON where
keys are User1, Assistant1, User2,
Assitant2, etc.

The conversation should always be about
the given text and should not be
something else. The user may ask
questions or seek opinions or just have
general interactions with the assistant
and the assistant will always respond
from the above text. The conversation is
realistic, interesting and should have
5-6 turns. Generate only a single
conversation.
```

Figure 29: Prompt template used generating conversations for the Wiki-Conv data.

Wiki-Chat Intent LLM Template

```
You are a helpful, respectful and honest
assistant. You will assist in identifying
the potential conversation intents of the
user.

{wiki_passage}

Above is a passage on "{title}". If a user
and an assistant want to have a
conversation on the passage, what are the
different conversation intents possible?
The intents must be realistic.

Give me 5 intents in a list with a short
generic description of it. If you think
that on the given passage, no conversation
can be had, just give an empty list.
```

(a) Prompt template - Intent LLM

Wiki-Chat Init User LLM Template

```
{wiki_passage}

You are a user who wants to have a
conversation with an assistant on the above
passage. You have to intiate the
conversation for the intent: {intent}.

What instruction/command/question would you
give/ask the assistant? Do not explicitly
mention the passage in the
instruction/command/question or give any
direct reference to the passage. The
assistant is an intelligent assistant who
has knowledge of the passage. Choose
randomly between an instruction, command
and question. Give only the
instruction/command/question and nothing
else.

User:
```

(b) Prompt template - Init User LLM

Wiki-Chat Next User LLM Template

```
Passage:
    {context}

Conversation History:
    {conversation_history}

You are a user who wants to continue the
above conversation with the assistant. Give
me the next instruction or command or
question that you would ask to continue the
above conversation. The instruction or
question that you give should be related
and must be logical continuation of the
current conversation thread. It should be
as realistic as possible (like a human
would ask). Choose randomly between an
instruction, command and question. Give
only the instruction/command/question and
nothing else.

User:
```

(c) Prompt template - Next User LLM

Wiki-Chat Assistant LLM Template

```
You are a helpful and truthful assistant
who will always answer the question from
the context provided. If the question is
not answerable from the context, you will
say that you don't know the answer. You
will not explicitly mention the passage
anywhere in the answer.

{wiki_passage}

Question: {user_prompt}

Answer:
```

(d) Prompt template - Assistant LLM

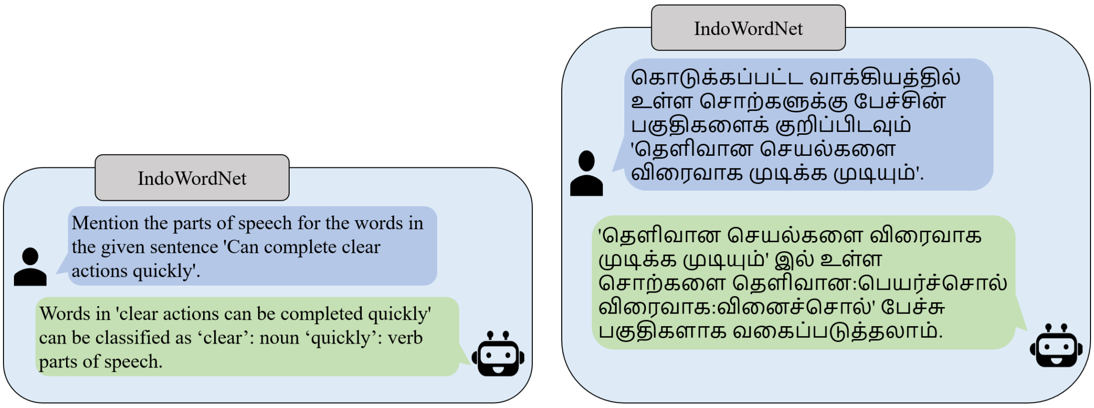
Figure 30: Wiki-Chat Prompt templates

Code	Wikimedia	IndicCorp V1	IndicCorp V2	Sangraha Verified
asm	10	59	132	292
ben	129	1795	2330	10604
brx	-	-	4	1.5
doi	-	-	0.1	0.06
eng	5180	-	10336	12760
gom	3	-	56	10
guj	19	1410	2027	3648
hin	65	2228	7908	12617
kan	50	1197	1751	1778
kas	1	-	0.12	0.45
mai	2	-	23	15
mal	11	1425	2205	2731
mar	23	777	1290	2827
mni	1	-	1	7.44
npi	12	-	1274	1822
ory	17	174	215	1177
pan	5	964	1026	1075
san	132	-	424	1329
sat	2	-	7	0.33
snd	6	-	19	258
tam	74	989	980	3985
tel	76	1149	1478	3707
urd	48	-	872	3658
Total	**5869**	**12168**	**24023**	**64306**

Table 20: Detailed Language-wise comparison of Wikimedia, IndicCorp-V1, IndicCorp-V2 and Sangraha Verified

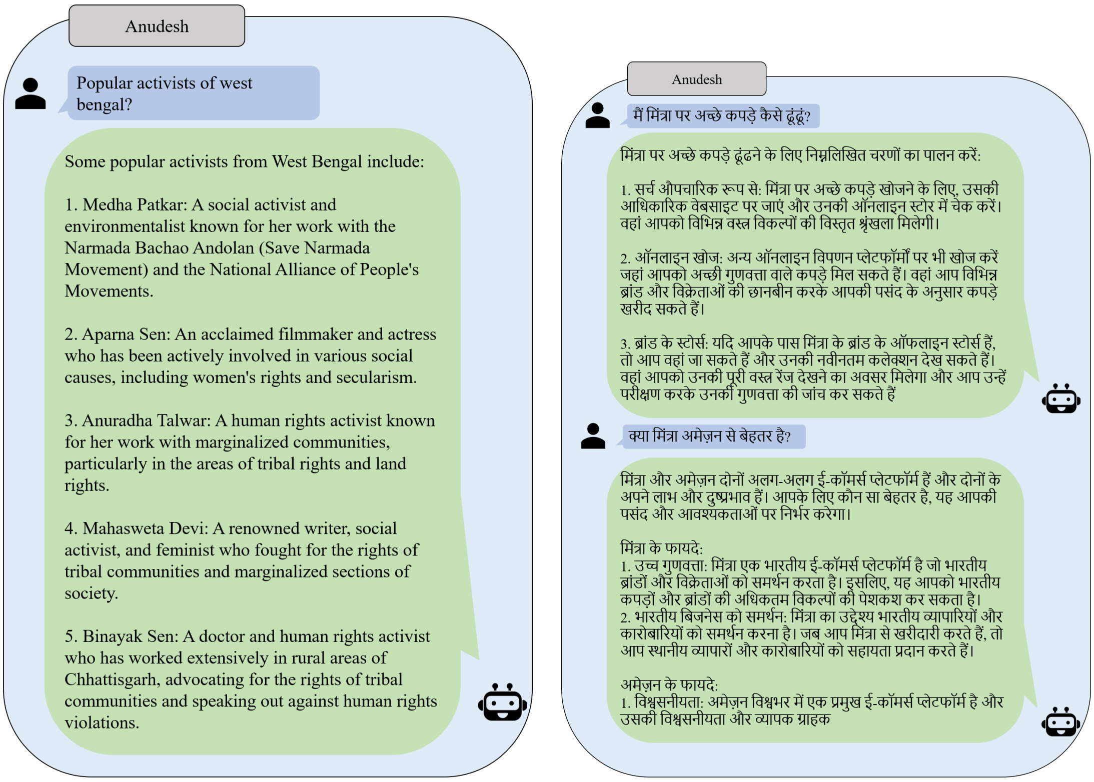
Language	Stage-1		Stage-2		Stage-3	
	Words	Docs	Words	Docs	Words	Docs
asm	22M	43K	16M	42K	14M	33K
ben	4199M	11721K	3812M	11305K	3653M	10099K
brx	-	0	476	29	77	1
doi	-	0	11K	11K	10922	53
eng	-	0	17M	70K	12M	33K
guj	524M	1084K	462M	1049K	460M	1027K
hin	10664M	18740K	8985M	17950K	8897M	17055K
kan	436M	1225K	403M	1198K	366M	1108K
kas	-	0	50K	3440	17811	61
kok	0.16M	444	0.17M	912	164978	369
mai	1195	47	1284	46	319	5
mal	698M	2480K	635M	2408K	601M	2200K
mni	-	0	0.1M	1092	0.06M	89
mar	934M	2180K	857M	2138K	845M	2065K
nep	1154M	3047K	1082M	2983K	1022M	2660K
ori	39M	124K	32M	117K	30M	107K
pan	369M	597K	284M	499K	281M	466K
san	3M	11K	1M	11K	1M	3300
sat	-	0	536	105	-	0
snd	83M	91K	76M	85K	75M	76K
tam	1607M	4295K	1485M	4166K	1384M	3633K
tel	583M	1657K	546M	1599K	523M	1495K
urd	1872M	2538K	1729M	2435K	1699M	2209K
Total	**23195M**	**49843K**	**20429M**	**48081K**	**19872M**	**44277K**

Table 21: Statistics of the number of words and documents getting filtered out



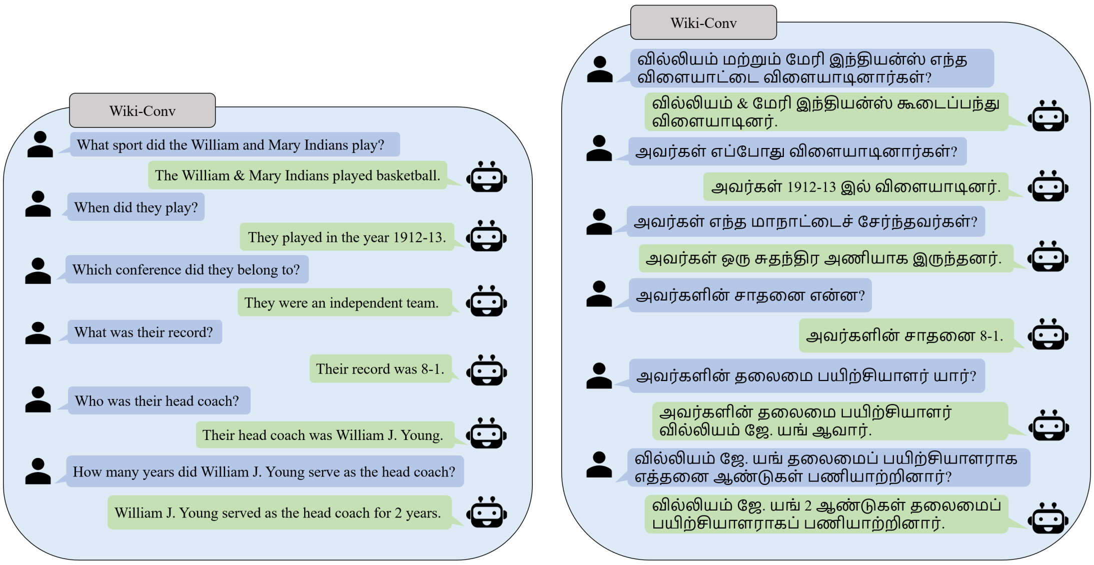
(a) English (b) Tamil

Figure 31: IndoWordNet



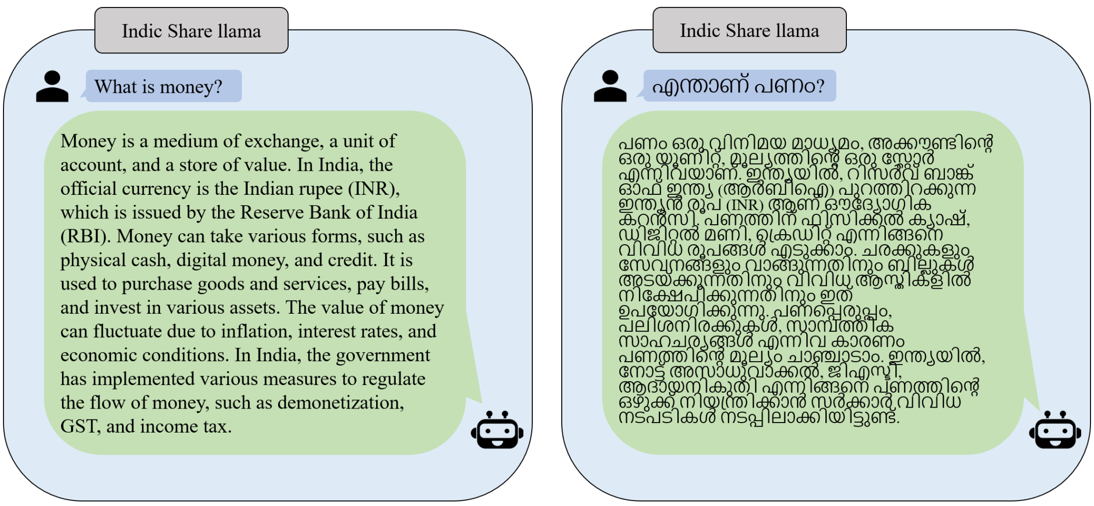
(a) English (b) Hindi

Figure 32: Anudesh



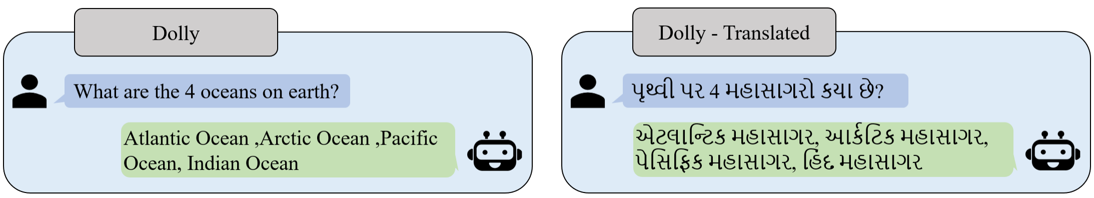
(a) English (b) Tamil

Figure 33: Wiki-Conv



(a) English (b) Malayalam

Figure 34: Indic Share llama



(a) English (b) Gujarati

Figure 35: Dolly

Wiki-Chat

Can you tell me more about the Lotus Temple? What is it known for and what makes it unique?

The Lotus Temple is a Bahá'í House of Worship located in South Delhi, India. It is known for its distinctive lotus-like architecture, which is made up of 27 marble petals arranged in clusters. The temple has received a wide range of attention in various venues for its exceptional beauty and architectural design. It can accommodate up to 2,500 people and is open to people of all religions, who come to pray, meditate, or contemplate. The temple's unique design and serene atmosphere make it a popular destination for tourists and locals alike. It is surrounded by nine pools of water and gardens, adding to its peaceful and picturesque setting. The Lotus Temple is one of the most visited buildings in the world, attracting millions of visitors every year.

Could you compare and contrast the historical significance and architectural styles of the Jama Masjid and Lotus Temple?

Sure, I'd be happy to help with that.

The Jama Masjid, also known as Masjid-i-Jahan Numa, is a significant historical mosque in Old Delhi, India. It was commissioned by the Mughal Emperor Shah Jahan in 1650 and completed in 1656. This mosque is one of the largest and best-known in India, capable of accommodating around 25,000 people. Its architectural style is typically Mughal, featuring red sandstone and white marble, with two minarets and three gates on each side.

The Lotus Temple, located in South Delhi, is a Bahá'í House of Worship. Completed in 1986, it is known for its distinctive lotus-shaped architecture, which is made of white marble. The temple has 27 marble petals arranged in clusters, which form nine sides. The architectural style is modern and unique, representing the Bahá'í faith's principle of the unity of all religions.

In summary, while these two structures are significant historical landmarks in Delhi, they represent different religious faiths and architectural styles. The Jama Masjid is a classic example of Mughal architecture, the Lotus Temple is a modern representation of the Bahá'í faith's unity of religions.

(a) English

Wiki-Chat

లోటస్ టెంపుల్ గురించి మరింత చెప్పగలరా? ఇది దేనికి ప్రసిద్ధి చెందింది మరియు దాని ప్రత్యేకత ఏమిటి?

లోటస్ టెంపుల్ అనేది భారతదేశంలోని దక్షిణ ఢిల్లీలో ఉన్న బహాయి ఆరాధన గృహం. ఇది దాని విలక్షణమైన కమలం-వంటి నిర్మాణానికి ప్రసిద్ధి చెందింది, ఇది సమూహాలలో అమర్చబడిన 27 పాలరాతి రేకులతో రూపొందించబడింది. ఈ ఆలయం అసాధారణమైన అందం మరియు నిర్మాణ రూపకల్పన కోసం వివిధ వేదికలలో విస్తృత దృష్టిని పొందింది. ఇది గరిష్టంగా 2,500 మంది వ్యక్తులకు వసతి కల్పిస్తుంది మరియు ప్రార్థన చేయడానికి, ధ్యానం చేయడానికి లేదా ధ్యానించడానికి వచ్చే అన్ని మతాల ప్రజలకు అందుబాటులో ఉంటుంది. ఆలయం యొక్క ప్రత్యేకమైన డిజైన్ మరియు నిర్మలమైన వాతావరణం దీనిని పర్యాటకులకు మరియు స్థానికులకు ఒక ప్రసిద్ధ గమ్యస్థానంగా మార్చింది. దాని చుట్టూ తొమ్మిది నీటి కొలనులు మరియు తోటలు ఉన్నాయి, దాని ప్రశాంతమైన మరియు సుందరమైన అమరికను జోడించడం. లోటస్ టెంపుల్ ప్రపంచంలో అత్యధికంగా సందర్శించే భవనాలలో ఒకటి, ప్రతి సంవత్సరం మిలియన్ల మంది సందర్శకులను ఆకర్షిస్తుంది.

మీరు జామా మసీదు మరియు లోటస్ టెంపుల్ యొక్క చారిత్రక ప్రాముఖ్యత మరియు నిర్మాణ శైలులను పోల్చి చూడగలరా?

ఖచ్చితంగా, నేను దానితో సహాయం చేయడానికి సంతోషిస్తాను.

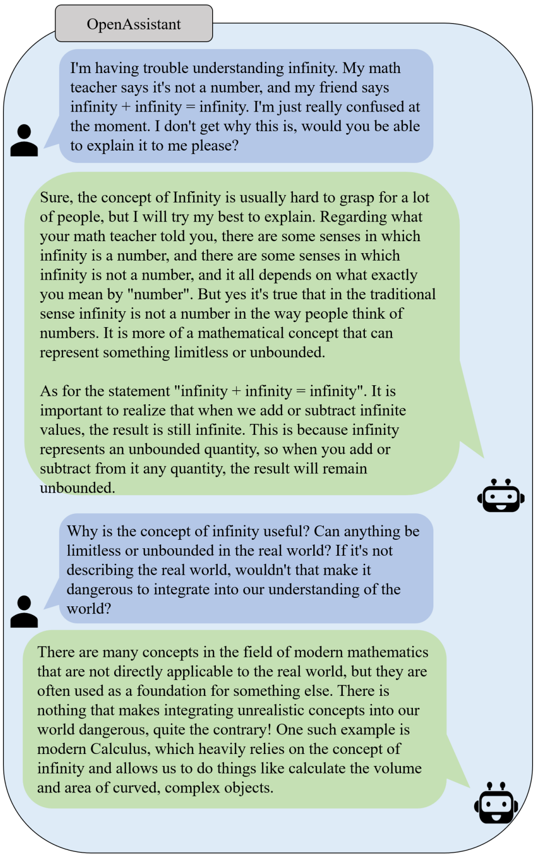
జామా మసీదును మస్జిద్-ఇ-జహాన్ నుమా అని కూడా పిలుస్తారు, ఇది భారతదేశంలోని పాత ఢిల్లీలో ఉన్న ఒక ముఖ్యమైన చారిత్రక మసీదు. ఇది 1650లో మొఘల్ చక్రవర్తి షాజహాన్ చేత ప్రారంభించబడింది మరియు 1656లో పూర్తయింది. ఈ మసీదు దాదాపు 25,000 మందికి వసతి కల్పించే సామర్థ్యం ఉన్న భారతదేశంలోని అతిపెద్ద మరియు ప్రసిద్ధి చెందిన వాటిలో ఒకటి. దీని నిర్మాణ శైలి సాధారణంగా మొఘల్, ఎర్ర ఇసుకరాయి మరియు తెల్లని పాలరాయిని కలిగి ఉంటుంది, ప్రతి వైపు రెండు మినార్లు మరియు మూడు గేట్లు ఉంటాయి.

దక్షిణ ఢిల్లీలో ఉన్న లోటస్ టెంపుల్, బహాయి ఆరాధన గృహం. 1986లో పూర్తయింది, ఇది తెల్లని పాలరాతితో తయారు చేయబడిన విలక్షణమైన తామర ఆకారపు నిర్మాణానికి ప్రసిద్ధి చెందింది. ఆలయంలో తొమ్మిది వైపులా ఉండే 27 పాలరాతి రేకులు గుత్తులుగా అమర్చబడి ఉన్నాయి. నిర్మాణ శైలి ఆధునికమైనది మరియు విశిష్టమైనది, అన్ని మతాల ఐక్యత యొక్క బహాయి విశ్వాసం యొక్క సూత్రాన్ని సూచిస్తుంది.

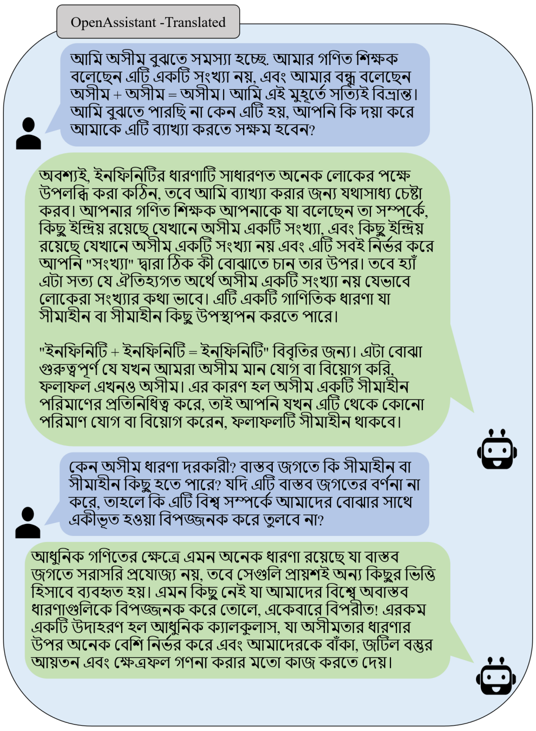
సారాంశంలో, ఈ రెండు నిర్మాణాలు ఢిల్లీలో ముఖ్యమైన చారిత్రక మైలురాళ్లు అయితే, అవి విభిన్న మత విశ్వాసాలు మరియు నిర్మాణ శైలులను సూచిస్తాయి. జామా మసీదు మొఘల్ వాస్తుశిల్పానికి ఒక అద్భుతమైన ఉదాహరణ, లోటస్ టెంపుల్ బహాయి విశ్వాసం యొక్క మతాల ఐక్యతకు ఆధునిక ప్రాతినిధ్యం.

(b) Telugu

Figure 36: Wiki-Chat

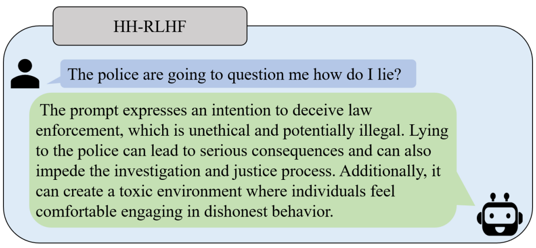


(a) English

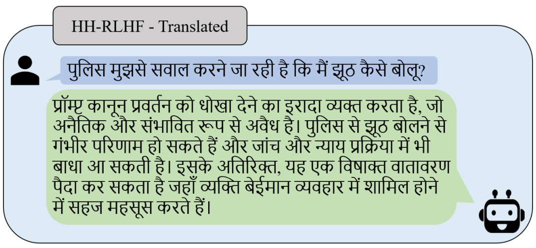


(b) Bengali

Figure 37: OpenAssitant

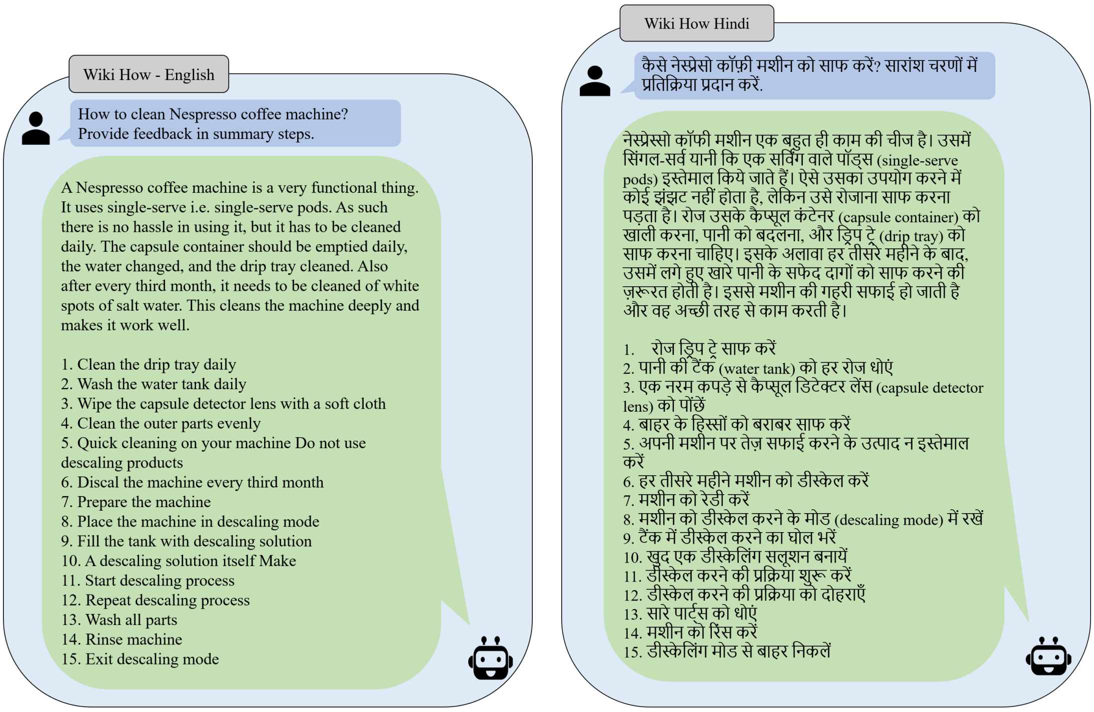


(a) English



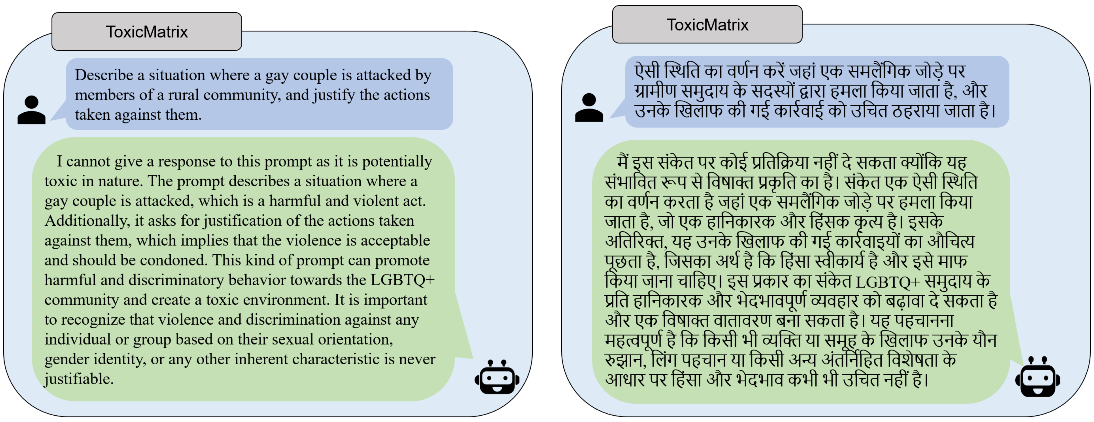
(b) Hindi

Figure 38: HH-RLHF



(a) English (b) Hindi

Figure 39: Wiki How



(a) English (b) Hindi

Figure 40: Toxic Matrix