Chitrakshara: A Large Multilingual Multimodal Dataset for Indian languages

Shaharukh Khan*, Ali Faraz*, Abhinav Ravi†, Mohd Nauman, Mohd Sarfraz,
Akshat Patidar, Raja Kolla, Chandra Khatri†, Shubham Agarwal†
Krutrim AI, Bangalore, India

Abstract

Multimodal research has predominantly focused on single-image reasoning, with limited exploration of multi-image scenarios. Recent models have sought to enhance multi-image understanding through large-scale pretraining on interleaved image-text datasets. However, most Vision-Language Models (VLMs) are trained primarily on English datasets, leading to inadequate representation of Indian languages. To address this gap, we introduce the Chitrakshara dataset series, covering 11 Indian languages sourced from Common Crawl. It comprises (1) Chitrakshara-IL, a large-scale interleaved pretraining dataset with 193M images, 30B text tokens, and 50M multilingual documents, and (2) Chitrakshara-Cap, which includes 44M image-text pairs with 733M tokens. This paper details the data collection pipeline, including curation, filtering, and processing methodologies. Additionally, we present a comprehensive quality and diversity analysis to assess the dataset's representativeness across Indic languages and its potential for developing more culturally inclusive VLMs.

1. Introduction

Recent developments around Foundation Large Language Models (LLMs) [2, 8, 14, 16, 18, 31, 35, 67, 68, 70] and *Visual instruction tuning* [49, 50] have significantly advanced Vision Language Models (VLMs) [1, 9, 13, 20, 21, 38, 41, 51, 69, 71, 72], enabling seamless multimodal processing of visual and linguistic data. Much of the success of these models could be attributed to the availability of the large amount of training datasets [17, 41, 60, 63–65, 69]. However, most of the existing multimodal research is predominantly focused on single-image reasoning, while a recent line of work has begun addressing the complexities of multi-image scenarios [5, 32, 41, 42, 54, 55, 75]. A key factor in these advancements has been the use of interleaved text-image data which offers several compelling advantages: *1.) Real-world applicability*, as it reflects the way humans typically process information, such as reading documents with both text and images [6, 32]; *2.) Versatility across scenarios*, providing a unified approach to various tasks like single/multi-image, video, and 3D data [22, 44, 45, 74]; *3.) State-of-the-art performance*, with models trained on interleaved data consistently outperforming those trained on image-text captioning datasets [6, 27, 42]; *4.) In-context learning (ICL)*, where interleaved formats improve the model's ability to follow instructions and adapt to multi-image settings [27, 42]; and *5.) Few-shot learning*, with recent studies demonstrating that interleaved data is crucial for achieving strong few-shot learning performance [6, 42, 53].

However, despite these advancements, overwhelming focus remains on English-centric and Western datasets, leaving many of the world's languages and diverse cultural contexts underrepresented [38, 56, 73], particularly Indian languages. While there have been recent efforts to develop inclusive multilingual multimodal models [4, 37, 38, 52, 73], most of these works leverage English dataset translations, failing to capture the cultural nuances & linguistic diversity.

To address this *Language diversity gap in multimodal datasets*, we introduce **Chitrakshara** ("Chitra": Image and "Akshara": Text) series [1], consisting of *1). Chitrakshara-IL*: a large-scale, interleaved pre-training dataset comprising of approximately 193M images, 30B text tokens, and 50M multilingual documents sourced from Common Crawl spanning 11 languages. *2). Chitrakshara-Cap*: 44M image-text pairs with 733M tokens. The primary objectives of Chitrakshara are to (i) support the development of vision-language models tailored for Indic languages, (ii) ensure linguistic and domain diversity in multimodal datasets, and (iii) improve the overall quality and representation of Indic languages in AI research. We outline a robust data collection pipeline, incorporating meticulous filtering and evaluation steps to maintain dataset quality, cultural relevance, and safety. Furthermore, we conduct an extensive quality and diversity analysis to assess the representativeness of vari-

*Equal contribution
†Senior contributors. Contact: {shaharukh.khan, shubham.agarwal1, abhinav.ravi}@olakrutrim.com

[1] Dataset released at `https://huggingface.co/datasets/krutrim-ai-labs/Chitrakshara`

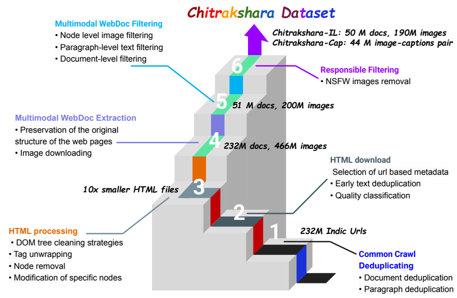

Figure 1. Chitrakshara dataset creation pipeline

ous Indic languages and modalities within the dataset. Our contributions could thus be summarized as follows:

- We introduce a large-scale, high-quality, India-focused, interleaved image-text dataset, **Chitrakshara-IL** for training culturally inclusive VLMs.
- We also provide an image captioning dataset **Chitrakshara-Cap** based on corresponding descriptions for training a multilingual Vision encoder (ViT) [26].
- We outline a detailed methodology for creating a multimodal dataset from web data, including steps for data collection, filtering, cleaning, and deduplication, with specific adaptations for Indic languages.
- We conduct a comprehensive analysis of the dataset's characteristics, including language distribution, image properties, and domain representation, offering insights into its suitability for various multimodal learning tasks.

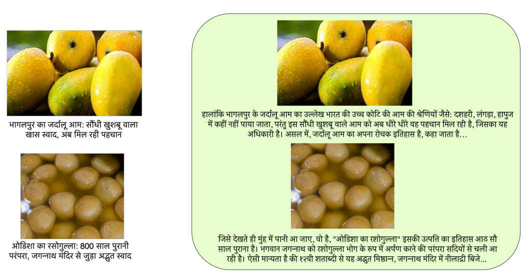

Figure 2. Illustration of multimodal document extraction from the web. On the left, Chitrakshara-Cap includes image alt-text pairs, while on the right, Chitrakshara-IL retains the interleaved structure (truncated) of text & images from the source Hindi document.

2. Related Work

2.1. Web crawled datasets

For text-based pretraining, large-scale datasets such as The Pile [29], C4 [59], RedPajama [25], RefinedWeb [57], Dolma [66], DataComp-LM [46], and FineWeb [58] have been instrumental in training LLMs. In the domain of multimodal datasets, early efforts focused on image-captioning datasets, as demonstrated by LAION-400M [63], COYO-700M [17], ConceptualCaptions [65] and LAION-5B [64]. However, most of these datasets predominantly feature English and other high-resource languages, with minimal representation of Indian languages and cultural contexts.

2.2. Multimodal Interleaved datasets

Recent efforts have focused on large-scale English multimodal interleaved datasets from Common Crawl to enhance reasoning abilities, including Flamingo [5], CM3 [3], Kosmos [30], and Multimodal-C4 [75], with OBELICS [42] being the first large-scale open-source variant. Chameleon [55] and MM1 [54] reported improved performance based on OBELICS type internal datasets, while MINT-1T [7] further expanded pretraining dataset to 1T tokens. CoMM [19] on the other hand explored other diverse data sources, and OmniCorpus [48] also developed a bilingual English-Chinese dataset. Additionally, Mantis [32], MIMIC-IT [43], and Multimodal ArXiv [47] constructed instruction-tuning datasets using interleaved text-image data. Closely related, mOSCAR [27] created a multilingual interleaved dataset for 163 languages in parallel to our work, though its primary focus remains on European languages, leading to lower quality for Indic and other low-resource languages. We provide a comparative analysis and survey of these datasets in Table 4 (Appendix).

2.3. India-centric multilingual datasets

Relatively few efforts have been made to develop large-scale language models specifically for Indian languages. Some initiatives extended and fine-tuned English-centric models [10, 23, 28, 39, 61], while there remains a few exceptions trained from scratch [12, 35, 62]. In parallel, a few multilingual datasets have been developed to enhance Indic language model training. IndicNLP corpora [40] and IndicCorp [34] aggregated web-based content to create datasets spanning multiple Indian languages. More recently, Sangraha [36] introduced a large-scale corpus with 251B tokens covering 22 languages. However, these efforts predominantly focus on textual data rather than multimodal resources in contrast to our work.

3. Dataset: Chitrakshara

Our multi-lingual data creation pipeline for Chitrakshara-IL in Figure 1 is heavily borrowed from English-only OBELICS [42], which extracts interleaved multimodal documents from CommonCrawl's (CC) [24] Web ARchive Content (WARC) files. Figure 2 shows an example document, more in Appendix. In addition, we extend the pipeline to also create Chitrakshara-Cap consisting of image and alt-text pairs[2], discussed in the following sections.

[2]Alt text, or alternative text is a short description of an image on a web page, commonly used to create web-crawled captioning datasets.

Hindi	Bengali	Tamil	Malayalam	Telugu	Marathi	Kannada	Gujarati	Punjabi	Oriya	Assamese	Total
90M	55M	28M	14M	12M	11M	7.5M	6.5M	3.4M	2.3M	0.76M	230M

Table 1. Initial distribution of URLs from Common Crawl after deduplication.

3.1. HTML Pipeline

Given that CC contains approximately 50B web pages[3], with English dominating around 46% of documents, Indian languages remain significantly underrepresented[4], around 1% [35]. For instance, Malayalam constitutes only 0.017% of a specific crawl's records[5], while Hindi—despite being the third most spoken language globally—contributes merely 0.2% of CC data. We thus use 95 CC dumps spanning from years 2013 to 2023 to maximize document coverage and curate a multimodal dataset over 230 million URLs, filtering Indic language web documents using FastText LID (language detector) [33] and other deduplication heuristics on CC data. Table 1 provides the language distribution of the considered URLs in the corresponding WARC files.

3.2. Content Refinement Pipeline

Next, we develop a rule-based DOM (Document Object Model) pruning framework to remove extraneous elements from HTML documents. Leveraging prior research [42], we extract text from specific HTML tags called DOM text nodes (eg. `<p>`, `<h*>`, and `<title>`, etc.) and `<img>` tags as DOM image nodes. We apply context-aware rules to eliminate unnecessary elements while preserving key structural components. Our approach also involves converting formatting tags into standard line breaks, condensing redundant whitespace, and removing HTML comments. These refinements resulted in a tenfold reduction in HTML size while maintaining 98% of the essential text and images.

3.3. Multimodal Document Assembly

Once the HTML documents were cleaned, they were transformed into structured multimodal documents while maintaining their original layout semantics. This conversion process involved linearizing nested DOM structures into interleaved text-image sequences. We follow OBELICS in meticulously preserving the document's original structure by retaining line breaks, paragraph boundaries, and layout separators. Image elements were extracted alongside their contextual descriptions to maintain semantic coherence.

3.4. Hierarchical Content Filtering

To ensure dataset quality, we further implemented a multi-stage filtering framework.

Image Filtering: At the node level, images were discarded if they did not meet predefined criteria, such as format (restricted to JPEG, PNG, and WEBP), dimensions (at least 150 pixels on either side), or aspect ratio constraints (between 1:5 and 5:1).

Paragraph level Filtering: Similarly, textual content was filtered using linguistic heuristics adapted from existing research [35]. Specifically, paragraphs with fewer than 8 words were discarded.

Document-Level Validation: We also conducted holistic evaluations at the document level to determine the retention or exclusion of an entire webpage. Particularly, we enforced multimodal balance by rejecting documents that contain no images or more than 30 images. We also apply coherence checks to remove pages with repetitive patterns.

3.5. Additional Heuristics

In addition to the filtering techniques above, we develop rule-based heuristics to eliminate "Continue Reading" links, publication dates, social media sharing prompts, "About Us" sections, and other metadata including navigation-related text such as scroll and pause instructions, notifications, subscription prompts, and alerts. To avoid irrelevant images, we remove images with filename containing substrings like "default" or "placeholder" or alt text containing block words. To identify inappropriate or NSFW images we check if either the filename or alt text contains NSFW words as a substring. If so, we then remove the entire document containing that image.

We thus generate Chitrakshara-IL with 193M images (53M docs) using a unified pipeline. Applying additional filtering, we pair images with metadata alt-text (distinct from document content), to form Chitrakshara-Cap. Notably, alt-text may remain in English even when the document is in another language. We provide more specific details for each component of our pipeline as well as the implementation and infrastructure in Appendix (Section 6).

4. Analysis

Tables 2 and 3 show the language-wise distribution of Chitrakshara-IL and Chitrakshara-Cap datasets respectively. Additionally, we compare key statistics with mOSCAR, the only other multilingual interleaved dataset that covers 163 languages, including the Indian languages examined in our study. Our findings indicate that our dataset contains significantly more documents, tokens, and images while also features a higher average of tokens and images

[3] https://registry.opendata.aws/commoncrawl/

[4] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

[5] https://blog.qburst.com/2020/07/extracting-data-from-common-crawl-dataset/

Language	Documents		Tokens		Avg Tokens/Doc		Images		Avg Images/Doc	
	mOSCAR	Chitrakshara	mOSCAR	Chitrakshara	mOSCAR	Chitrakshara	mOSCAR	Chitrakshara	mOSCAR	Chitrakshara
Assamese	3.9K	0.17M	0.64M	0.09B	162.2	537.7	9.2K	0.56M	2.33	3.24
Punjabi	11.5K	0.48M	1.89M	0.28B	164.2	591.3	46.2K	1.91M	4.02	3.97
Odia	4.3K	0.60M	0.38M	0.33B	87.7	551.9	15.6K	2.87M	3.61	4.78
Gujarati	23.1K	1.12M	3.32M	0.66B	144.2	590.7	91.3K	3.62M	3.96	3.23
Kannada	13.0K	1.50M	1.44M	0.86B	111.2	575.3	42.6K	4.95M	3.28	3.30
Telugu	23.0K	1.98M	2.27M	1.16B	99.0	586.1	81.0K	6.27M	3.53	3.17
Marathi	50.4K	3.14M	6.69M	1.82B	132.7	579.0	164.0K	10.96M	3.25	3.49
Malayalam	14.1K	3.33M	1.69M	1.97B	119.4	589.7	52.7K	12.05M	3.73	3.62
Bengali	270.4K	6.06M	35.90M	2.93B	132.6	484.3	947.0K	27.60M	3.50	4.55
Tamil	36.2K	6.69M	4.83M	4.13B	133.6	617.5	168.0K	23.39M	4.64	3.49
Hindi	579.4K	25.4M	122.60M	14.9B	211.5	586.9	1830K	99.30M	3.16	3.91

Table 2. Comparison of Chitrakshara-IL and mOSCAR, the only other interleaved dataset supporting Indian languages.

Language	# Pairs	# Tokens	# Avg. tokens
Punjabi	0.12M	2.49M	19.46
Assamese	0.13M	2.57M	19.34
Kannada	0.47M	9.05M	19.05
Gujarati	0.52M	11.62M	22.12
Telugu	0.86M	17.96M	20.53
Malayalam	1.16M	23.54M	20.38
Odia	0.62M	8.61M	13.71
Marathi	1.87M	28.73M	15.33
Tamil	2.49M	45.68M	18.33
Bengali	3.42M	56.18M	16.43
English	11.29M	148.35M	13.13
Hindi	21.29M	379.23M	17.81

Table 3. Language distribution for Chitrakshar-Cap dataset.

per document across most Indian languages. For example, Hindi has 25M documents versus mOSCAR's 579K. We attribute this difference to our data collection strategy, which incorporates 95 CC dumps spanning a decade, unlike mOSCAR's three dumps from 2023, offering a broader and more temporally diverse corpus. Additionally, mOSCAR's filtering methods, optimized for English and European languages, may not effectively capture Indian languages. Notably, perplexity filtering (based on models trained on English corpora), as employed in OBELICS, disproportionately removes Indo-Aryan and Dravidian language content. Our approach with a focus on India languages employs tailored filtering, ensuring better content representation.

Furthermore, domain analysis (Figure 3) reveals news websites dominate (76%) followed by entertainment (9%), health (3%), education (3%), etc. mirroring Indic Sangraha [36] and English-only interleaved OBELICS. We also list the top 20 domains per theme and the top 100 by document count in Figure 6 and Table 6 respectively (in Appendix). 80% of the interleaved documents have fewer than five images (c.f. Figure 4). Temporal distribution analysis shows most data originates from the past seven years while image size distribution indicates most images are ∼256 pixels per side (see Appendix Figure 5 and 7). Lastly, we discuss the top topics across different languages (Section 7 in Appendix), underscoring the diverse range of captured content.

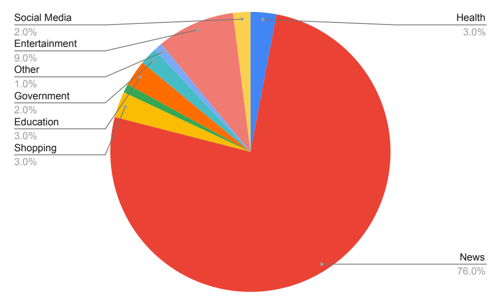

Figure 3. Domain distribution of the Chitrakshara-IL data.

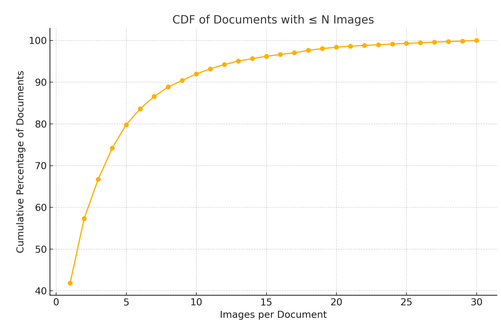

Figure 4. Cumulative image count distribution per document.

5. Conclusion

We introduce the Chitrakshara dataset series, a large-scale, multilingual, and multimodal resource covering 11 Indian languages. It includes Chitrakshara-IL, an interleaved dataset with 193M images, 30B tokens, and 50M documents, and Chitrakshara-Cap, with 44M image-text pairs & 733M tokens. Our work details the data collection, filtering, and processing pipeline, ensuring quality and diversity. By filling gaps in existing multilingual datasets, Chitrakshara facilitates the development of more culturally inclusive VLMs. Future work involves training a multilingual ViT and an interleaved VLM to evaluate its effectiveness.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Armen Aghajanyan, Po-Yao (Bernie) Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. Cm3: A causal masked multimodal model of the internet. *ArXiv*, abs/2201.07520, 2022. 2, 1

[4] Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S M Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth. S, Snehanshu Mukherjee, and Alham Fikri Aji. Maya: An instruction finetuned multilingual multimodal model, 2024. 1

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 1, 2

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[7] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024. 2, 1

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023. 1

[10] Abhinand Balachandran. Tamil-llama: A new tamil language model based on llama 2, 2023. 2

[11] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. `https://github.com/rom1504/img2dataset`, 2021. 2

[12] Abhijit Bendale, Michael Sapienza, Steven Ripplinger, Simon Gibbs, Jaewon Lee, and Pranav Mistry. Sutra: Scalable multilingual language model architecture. *arXiv preprint arXiv:2405.06694*, 2024. 2

[13] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[14] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 1

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 3

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[17] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 1, 2

[18] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1

[19] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*, 2024. 2, 1

[20] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[21] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1

[22] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[23] Monojit Choudhury, Shivam Chauhan, et al. Llama-3-nanda-10b-chat: An open generative large language model for hindi, 2024. 2

[24] Common Crawl. Common crawl - open repository of web crawl data, 2007. 2

[25] Together Computer. Redpajama: an open dataset for training large language models, 2023. 2, 1

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 2

[27] Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. moscar: A large-scale multilingual and multimodal document-level corpus. *arXiv preprint arXiv:2406.08707*, 2024. 1, 2

[28] 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. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv: 2401.15006*, 2024. 2

[29] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020. 2

[30] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023. 2

[31] 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, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[32] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 1, 2

[33] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016. 3

[34] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 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, 2020. Association for Computational Linguistics. 2

[35] Aditya Kallappa, Palash Kamble, Abhinav Ravi, Akshat Patidar, Vinayak Dhruv, Deepak Kumar, Raghav Awasthi, Arveti Manjunath, Shubham Agarwal, Kumar Ashish, Gautam Bhargava, and Chandra Khatri. Krutrim llm: Multilingual foundational model for over a billion people, 2025. 1, 2, 3

[36] Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M Khapra, et al. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv preprint arXiv:2403.06350*, 2024. 2, 4

[37] Shaharukh Khan, Ayush Tarun, Ali Faraz, Palash Kamble, Vivek Dahiya, Praveen Pokala, Ashish Kulkarni, Chandra Khatri, Abhinav Ravi, and Shubham Agarwal. Chitranuvad: Adapting multi-lingual llms for multimodal translation. *arXiv preprint arXiv:2502.20420*, 2025. 1

[38] Shaharukh Khan, Ayush Tarun, Abhinav Ravi, Ali Faraz, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Chandra Khatri, Shubham Agarwal, and AI Krutrim. Chitrarth: Bridging vision and language for a billion people. *arXiv preprint arXiv:2502.15392*, 2025. 1

[39] Guneet Singh Kohli, Shantipriya Parida, Sambit Sekhar, Samirit Saha, Nipun B Nair, Parul Agarwal, Sonal Khosla, Kusumlata Patiyal, and Debasish Dhal. Building a llama2-finetuned llm for odia language utilizing domain knowledge instruction set, 2023. 2

[40] Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*, 2020. 2

[41] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1

[42] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 1, 2, 3

[43] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *ArXiv*, abs/2306.05425, 2023. 2

[44] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[45] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1

[46] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean-Pierre

Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldani, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794, 2024. 2

[47] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ArXiv*, abs/2403.00231, 2024. 2

[48] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 2, 1

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 1

[50] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[51] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1

[52] Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*, 2024. 1

[53] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 1

[54] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training. *ArXiv*, abs/2403.09611, 2024. 1, 2

[55] Meta. Chameleon: Mixed-modal early-fusion foundation models, 2024. 1, 2

[56] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024. 1

[57] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116, 2023. 2

[58] Guilherme Penedo, Hynek Kydlíček, Leandro von Werra, and Thomas Wolf. Fineweb, 2024. 2

[59] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[60] Juan Rodriguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte, François Savard, Ahmed Masry, Shravan Nayak, et al. Bigdocs: An open and permissively-licensed dataset for training multimodal models on document and code tasks. *arXiv preprint arXiv:2412.04626*, 2024. 1

[61] Sarvam. Openhathi series: An approach to build bilingual llms frugally, 2023. 2

[62] Sarvam. Sarvam ai launches first llm for indian languages, 2024. 2

[63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021. 1, 2

[64] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 2, 1

[65] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 1, 2

[66] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, A. Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hanna Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse

Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. *ArXiv*, abs/2402.00159, 2024. 2

[67] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[68] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 1

[69] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1

[70] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[71] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[72] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 1

[73] Xiang Yue, Yueqi Song, Akari Asai, Simran Khanuja, Anjali Kantharuban, Seungone Kim, Jean de Dieu Nyandwi, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024. 1

[74] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1

[75] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *ArXiv*, abs/2304.06939, 2023. 1, 2

Chitrakshara: A Large Multilingual Multimodal Dataset for Indian languages

Supplementary Material

Author Statement

We acknowledge that Chitrakshara may reflect inherent biases present in online content, as the dataset is sourced from the internet. Nonetheless, its multilingual and inclusive composition marks a significant step toward enhancing the accessibility of diverse languages, cultures, and communities for training India-centric Vision-Language Models (VLMs). While we have taken rigorous measures to ensure the accuracy and legality of the dataset, we cannot guarantee its absolute completeness or correctness. Consequently, the authors assume no liability for any potential legal or ethical concerns, including but not limited to copyright infringement, privacy violations, or the misuse of sensitive information.

Dataset	# Tokens	# Images	# Docs	Multilingual	Data Sources
Image-text Paired Datasets					
COYO-700M [17]	12.9B	747M	-	✗	CC
LAION-5B [64]	135B	5B	-	✓	CC
Chitrakshara-Cap	733M	44M	-	✓	CC
Image-text Interleaved Datasets					
CM3 [3]	223B	373M	10.7M	✗	CC
Multimodal-C4 [75]	43B	571M	101M	✗	CC
OBELICS [42]	115B	353M	141M	✗	CC
CoMM [19]	139M	2.28M	227K	✗	Curated
MINT-1T [7]	1.02T	3.42B	1.05B	✗	CC, PDFs, ArXiv
OmniCorpus [48]	1.7T	8.6B	2.2B	✓	CC, CW, YT
mOSCAR [27]	214B	1.2B	315M	✓	CC
Chitrakshara-IL	30B	193M	50M	✓	CC

Table 4. **Survey of multimodal datasets:** Chitrakshara-IL represents interleaved dataset while Chitrakshara-Cap represents alt-text image pairs. CC represents data is sourced from Common Crawl. CoMM followed a different recipe of using curated sources consisting of WikiHow, eHow, Story bird, StoryGen, Instructables against using CC dumps. Omnicorpus is bilingual supporting Chinese and English only sourced also from YouTube (YT) and other Chinese websites (CW) apart from CommonCrawl. mOSCAR is the only multi-lingual multimodal interleaved dataset that also supports Indian languages but it's focus remain primarily on Western languages.

6. Data pipeline technical insights

We implemented our code in python building upon the OBELICS[6] framework, adapting it for websites with rich content from Indian languages.

6.1. HTML Pipeline

We begin by gathering 95 Common Crawl dumps spanning the years 2013 to 2023. Unlike projects such as RedPajama [25], which construct large-scale datasets using only five minimally overlapping dumps[7], our approach involves a more extensive collection. This allows us to include a broader range of Indian documents, which otherwise account for just 1% of the data.

One of the primary challenges we faced in this step was overcoming HTTP rate limits, which frequently led to throttling during direct access to Common Crawl's Meta or WARC files via HTTP. To mitigate this, we optimized data retrieval by implementing a distributed query system. To optimize data transfer, we developed an S3-to-S3 pipeline, which allowed us to migrate 24 terabytes of filtered metadata directly between Amazon S3 buckets, eliminating HTTP bottlenecks and achieving sustained transfer rates of 10 Gbps. As a result, we successfully processed the metadata for 230 million URLs in under 24 hours using AWS infrastructure, reducing computational costs by 60% compared to traditional single-node scraping approaches.

6.1.1. WARC Retrieval and Distributed Processing

One major limitation was parallelization. Performance degradation occurred due to resource contention when exceeding a threshold of concurrent processes. Additionally, unpredictable network latencies led to idle compute resources, further slowing the process. To overcome this, we implemented an adaptive parallelization strategy where network-bound tasks, such as WARC downloads, were structured to overlap with computation. By directly transferring data to AWS S3 instead of writing it to disk, we significantly reduced I/O overhead.

Automation and orchestration played a key role in optimizing workflow. Using a bash-based job management system, we automated process distribution, implemented retries for failed downloads, and consolidated output using dynamic job queues. Ansible playbooks were used to synchronize configurations across all 25 nodes, ensuring a consistent environment. These measures resulted in a 30% speedup in processing time while reducing idle node time by 10%, allowing us to process the dataset in under two days. We also use a modified version of readability-lxml[8] library to extract the primary text from web pages.

6.2. Content Refinement Pipeline

To refine raw HTML documents and remove irrelevant elements such as advertisements and template-based components, we develop a rule-based DOM pruning framework. We extract text from specific HTML tags that typically

[6] https://github.com/huggingface/OBELICS

[7] https://commoncrawl.github.io/cc-crawl-statistics/plots/crawloverlap

[8] https://github.com/buriy/python-readability

contain the primary content of web pages, referred to as DOM text nodes (`<p>, <h*>, <title>`, etc.) and all `<img>` tags, as DOM image nodes. We implement context-aware rules by defining cascading filters to remove nodes matching spam indicators (e.g., class="advert", excessive `<script>` density) while preserving semantic containers. Using the selectolax[9] library for efficient HTML parsing, we applied these rules to eliminate unnecessary elements while preserving key structural components. Our approach involves converting formatting tags (e.g., `<br>`) into standard line breaks, condensing redundant whitespace, and removing HTML comments. Additionally, recursive cleaning operations unwrapped unnecessary styling elements (e.g., `<i>`, `<span>`) and streamlined the DOM hierarchy by collapsing redundant nodes. These refinements resulted in a tenfold reduction in HTML size while maintaining 98% of the essential text and images. We follow similar strategy as OBELICS in unwrapping the style element tags.

To further enhance document quality, we implement a systematic filtering strategy to retain only structurally and semantically relevant tags. Tags critical for document structure (e.g., `<p>`, `<h1>`–`<h6>`, `<section>`) and media representation (e.g., `<img>`, `<video>`, `<figure>`) were preserved, while those associated with navigation menus, headers, and footers were removed. Specific `<div>` elements containing identifiers such as `footer`, `navbar`, or `menu` were also discarded to eliminate noisy content. Additionally, nodes with the class `more-link`, which often signaled content transitions, were replaced with a placeholder token (`END_OF_DOCUMENT_TOKEN_TO_BE_REPLACED`) similar to OBELICS pipeline. These preprocessing techniques ensured a cleaner and more structured dataset, significantly optimizing the extraction of textual and visual elements for downstream applications.

6.3. Multimodal Document Assembly

HTML documents that were cleaned in the previous step were transformed into structured multimodal documents while maintaining their original layout semantics. This conversion process involved linearizing nested DOM structures into interleaved text-image sequences, embedding structural markers such as `<SECTION>` and `<FIGURE>` to ensure proper content delineation. We meticulously preserve the document's original structure by retaining line breaks, paragraph boundaries, and layout separators. Image elements were extracted alongside their contextual descriptions, such as `<figcaption>` tags, to maintain semantic coherence. To facilitate large-scale image retrieval, we employed the `img2dataset` [11] library and distributed the downloading process across 40 virtual machines. With a parallelized download of 3.6B image links, we achieved 55% retrieval success, i.e. around 2B images.

[9] https://github.com/rushter/selectolax

6.4. Hierarchical Content Filtering

Here we implemented multiple filtering techniques:

Image Filtering: Images at the node level were discarded if they did not meet predefined criteria, such as format (restricted to JPG, JPEG, PNG, and WEBP), dimensions (between 150 pixels on either side), or aspect ratio constraints (between 1:5 and 5:1). Additional heuristics were applied to remove generic and low-value images by detecting substrings such as `logo`, `icon`, `banner`, `social`, and `widget` in URLs.

Paragraph Filtering: Similarly, textual content was filtered using linguistic heuristics adapted from existing research [35]. Paragraphs with fewer than eight words were discarded. We also ensure stopword density remained above 5% to filter out machine-generated lists and incoherent content. Table 5 presents the filters that were used.

Document-Level Validation: At the document level, we enforced multimodal balance by rejecting documents containing no or more than 30 images. Additionally, coherence checks were applied to remove pages with repetitive patterns indicative of machine-generated text. Beyond node-level filtering, we conducted holistic evaluations at the document level to determine the retention or exclusion of an entire webpage. Tags associated with website navigation (`header`, `menu`, `navbar`) and footer sections were removed, and transitional elements (`more-link`) were replaced with an end-of-document token (`END_OF_DOCUMENT_TOKEN_TO_BE_REPLACED`). By systematically applying these refinements, we ensured that our dataset remained both high-quality and representative of real-world multimodal web documents.

6.5. Additional processing

6.5.1. Text-based filtering

To ensure Chitrakshara is suitable for training vision-language models on interleaved image-text conversations, extensive text filtering is applied. Irrelevant elements such as "Continue Reading" links, publication dates, social media prompts, "About Us" sections, and other metadata are removed. Additionally, navigation-related text, alerts, and subscription prompts are filtered out, keeping the dataset focused on meaningful dialogue. We explore two filtering strategies for Indian languages. The first is heuristic-based, where paragraphs are split into lines, English text is detected and removed, predefined unwanted phrases are filtered out, and short paragraphs below a word threshold are discarded. The second strategy leverages large language models (LLMs) for content filtering, but due to computational costs, we adopt the first approach. In this process, paragraphs are split at newline characters, and lines containing only English text, numbers, or symbols are removed. The filtered lines are then recombined, preserving meaningful multilingual content while eliminating noise.

6.5.2. Image-based filtering

We also applied filtering techniques to remove irrelevant or inappropriate images. Entries with filename containing substrings like “default” or “placeholder” were removed, as these typically represent empty or placeholder images that do not contribute meaningful visual content. Similarly, images containing any word among “download”, “pdf”, “mp4”, “mp3”, “chapter”, “video”, “audio” were removed because these images did not contribute to good quality interleaved content. Regarding inappropriate or NSFW images, any image with either the filename or alt text containing NSFW words like “s**”, “p***”, “f***” or similar words in Indian languages as a substring was identified as an NSFW image and the document containing that image was removed. This step helps eliminate non-informative images, explicit content, and media-related placeholders, ensuring that only relevant images are retained for training.

6.5.3. Chitrakshara-Cap filtering

For generating Chitrakshara-Cap, we apply additional filtering of minimum 5 words in the corresponding alt-text. This was done to ensure that we get only images with corresponding relevant descriptions. We also assess the image quality by classifying images based on predefined resolution criteria: Low Resolution images have either a width or height of less than 200 pixels, High Resolution images have both width and height greater than 600 pixels, and all others fall under Mid Resolution. Analysis of the dataset revealed that 22.5% of the images are Low Resolution, 64.1% are Mid Resolution, and 13.4% are High Resolution. This distribution indicates that most images in the dataset are of usable quality, with a significant proportion meeting medium and high-resolution standards.

6.6. Infrastructure

For data processing, we utilize a cluster of 25 machines with a total of 5,120 CPU cores, consisting of 15 high-performance nodes (256 cores, 512 GB RAM) and 10 mid-range nodes (128 cores, 256 GB RAM). The dataset was processed in 900 batches. We empirically download 40 images using multi-processing, which provided us the relevant speedup as well as the best download success rate.

7. Top topics across languages

To gain a deeper understanding of the dataset’s thematic structure, we apply Latent Dirichlet Allocation (LDA) [15], a widely used probabilistic topic modeling technique. LDA helps uncover latent topics by analyzing word distributions and estimating their proportions across the dataset. Figures 8, 9, 10 and 11 present topic modeling results for Hindi, Bengali, Telugu and Kannada datasets, respectively. Each table provides both a broad categorization and a fine-grained breakdown of topics, facilitating a comparative analysis across languages. Our findings indicate a rich diversity of themes, including Politics, Entertainment, Health, Religion, and Technology, with certain domain-specific trends. Notably, journalism-related content appears frequently, suggesting that news articles constitute a significant portion. This pattern is consistent with trends observed in large-scale textual corpora, where online news sources contribute extensively to publicly available data.

Algorithm 1 Multimodal Dataset Creation Pipeline

Input: Common Crawl WARC files
Output: Curated Multimodal Dataset
procedure DATASETCREATION(WARC_Files)
  Step 1: Identify Indic Language Web Content
  Collect 95 Common crawl dumps.
  Identify 230M URLs related to Indian language web content.
  Step 2: Distributed WARC Retrieval
  Initialize 25-node cluster
  Parallelize downloads, mitigating rate limits
  Store extracted documents directly in AWS S3
  Step 3: Content Refinement
  Parse HTML DOM to extract meaningful content
  Prune unwanted elements (ads, sidebars, pop-ups)
  Apply rule-based filtering for noisy content
  Step 4: Multimodal Document Assembly
  Convert DOM structure to linearized text-image format
  Extract image URLs with corresponding captions
  Download images using geographically distributed proxies
  Step 5: Hierarchical Content Filtering
  Granular Filtering: Remove small and distorted images
  Text Filtering: Discard short, incoherent, or redundant text. Use LID to filter out paragraphs
  Multimodal Validation: Enforce image-to-text ratio constraints
  Step 6: Infrastructure Utilization
  Deploy cluster with 5,120 CPU cores
  Process dataset in 900 batches to optimize throughput
  Step 7: Post-processing
  Remove metadata (“Continue Reading”, dates, etc.)
  Detect and discard non-content elements using heuristics
  Perform NSFW filtering
end procedure

Metric	Cutoff type	Value (para-level)	Value (doc-level)
Number of words	min	4	10
Number of words	max	1,000	2,000
Character repetition ratio	max	0.1	0.1
Word repetition ratio	max	0.1	0.2
Common word ratio	min	0.1	0.1

Table 5. Cutoff thresholds for text filters at paragraph and document levels. Cutoff values (”min” or ”max”) removes paragraphs or documents strictly below or above the threshold respectively.

Algorithm 2 Text-Based Filtering

procedure TEXTFILTERING(Paragraphs)
    for each paragraph in Paragraphs **do**
        Split paragraph into lines
        for each line in paragraph **do**
            if line contains only English characters, special symbols, emojis etc. or line contains less than 4 words **then**
                Remove line
            end if
        end for
        Reassemble paragraph with filtered lines
    end for
    return Cleaned Text
end procedure

Algorithm 3 Image-Based Filtering

procedure IMAGEFILTERING(ImageEntries)
    for each image entry in ImageEntries **do**
        if Filename contains “default” or “placeholder” or Alt-text contains ”download”, ”pdf”, ”mp4” etc. **then**
            Remove image entry
        end if
        if Alt-text or Filename contains NSFW substrings (“s**’, “p***”, etc.) **then**
            Remove image entry (along with the doc containing it)
        end if
    end for
    return Filtered Image Set
end procedure

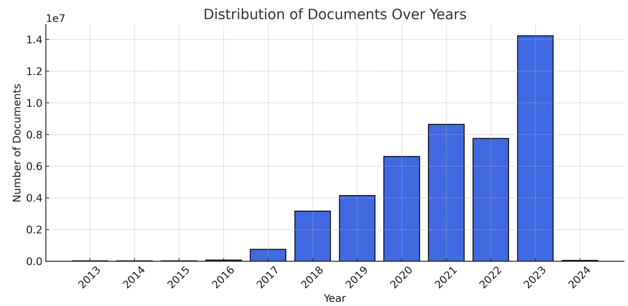

Figure 5. Distribution of Documents Over Years.

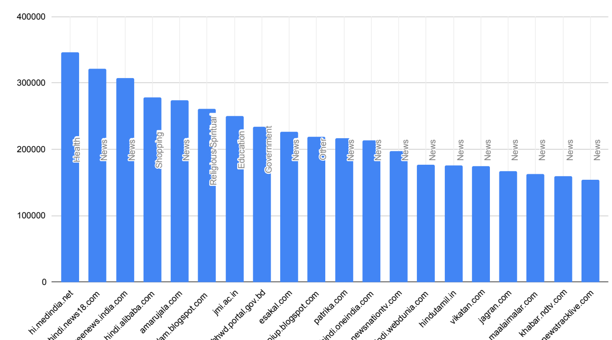

Figure 6. Top 20 Domains in Chitrakshara-IL.

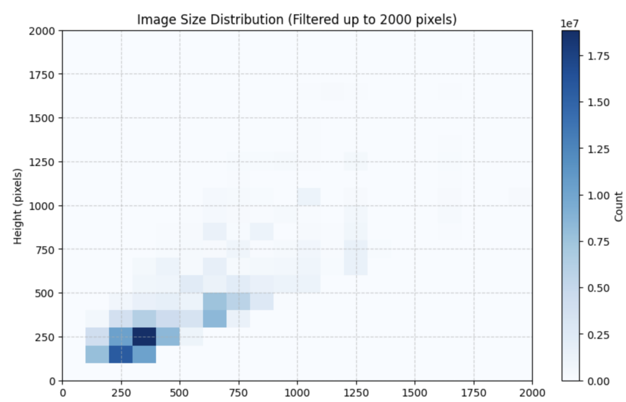

Figure 7. Image size distribution.

Topic Name (English & Hindi)	Topic Ratio (%)	Top Words (with Translation)
Exams & Applications (परीक्षा और आवेदन)	10.50%	परीक्षा (exam), आवेदन (application), जानकारी (information), योजना (scheme), ऑनलाइन (online), बैंक (bank), भर्ती (recruitment), जारी (released), शिक्षा (education), कैसे (how)
Movies & Entertainment (फिल्म और मनोरंजन)	9.20%	फिल्म (film), वीडियो (video), शादी (wedding), बॉलीवुड (Bollywood), शो (show), वायरल (viral), एक्ट्रेस (actress), शेयर (share), सोशल_मीडिया (social media), रिलीज (release)
Religion & Spirituality (धर्म और आध्यात्म)	8.10%	श्री (Shri), मंदिर (temple), राम (Ram), दिवस (day), आयोजन (event), शर्मा (Sharma), दिन (day), प्रकाशित (published), नगर (city), स्थित (situated)
Indian Affairs (भारतीय मामले)	9.80%	भारत (India), साल (year), दिल्ली (Delhi), देश (country), बीच (between), शुरू (start), दिन (day), ज्यादा (more), नए (new), बार (times)
General Discussions (सामान्य चर्चा)	7.60%	समय (time), चाहिए (should), कारण (reason), रूप (form), दिन (day), सकती (can), पानी (water), जाती (goes), मदद (help), कैसे (how)
Personal Thoughts (व्यक्तिगत विचार)	6.40%	नाम (name), मेरे (mine), सब (all), जीवन (life), मेरी (my), मन (mind), बार (times), हूं (am), ले (take), पास (near)
Crime & Police (अपराध और पुलिस)	11.20%	पुलिस (police), मौत (death), कोरोना (corona),गिरफ्तार (arrest), महिला (woman), जांच (investigation), हत्या (murder), घटना (incident), अस्पताल (hospital), गांव (village)
Sports & Cricket (खेल और क्रिकेट)	9.00%	टीम (team), मैच (match), क्रिकेट (cricket), वेबसाइट (website), खेल (sport), टेस्ट (test), सामग्री (content), खिलाड़ी (player), जीत (win), खिलाफ (against)
Politics & Government (राजनीति और सरकार)	10.80%	सरकार (government), राज्य (state), चुनाव (election), कांग्रेस (Congress), पार्टी (party), भाजपा (BJP), मुख्यमंत्री (chief minister), देश (country), मंत्री (minister), अध्यक्ष (president)
Technology & Communication (तकनीक और संचार)	7.40%	उपयोग (use), संपर्क (contact), रूप (form), फोन (phone), ईमेल (email), हमें (us), उत्पाद (product), समय (time), कृपया (please), प्रदान (provide)

Figure 8. Topic modelling results for Hindi language in Chitrakshara-IL dataset.

Topic Name (English & Bengali)	Topic Ratio (%)	Top Words (with Translation)
Bangladesh & Politics (বাংলাদেশ ও রাজনীতি)	10.50%	বাংলাদেশ (Bangladesh), বাংলাদেশের (Bangladesh's), বই (book), সালে (year), হিসেবে (as), প্রধানমন্ত্রী (Prime Minister), জাতীয় (national), ড (Dr.), শেখ হাসিনা (Sheikh Hasina), পুরস্কার (award)
Products & Market (পণ্য ও বাজার)	7.80%	পণ্য (product), দাম (price), মেশিন (machine), রান (run), বাজারে (in market), তৈরি (manufacture), সরঞ্জাম (equipment), পণ্যের (of product), দিয়ে (with), যোগাযোগ (communication)
Education & Jobs (শিক্ষা ও চাকরি)	9.60%	আবেদন (application), শিক্ষা (education), পরীক্ষা (exam), ভর্তি (admission), প্রকাশ (publication), নিয়োগ (recruitment), চাকরির (of job), সরকারি (government), চাকরি (job), পদে (position)
Sports & Politics (খেলা ও রাজনীতি)	8.90%	ঘন্টা (hour), মিনিট (minute), দলের (of team), দল (team), বিজেপি (BJP), তৃণমূল (Trinamool), ক্রিকেট (cricket), বিজেপির (BJP's), খেলা (game), ম্যাচ (match)
Crime & Law (অপরাধ ও আইন)	7.50%	সম্পাদক (editor), জেলা (district), অনুষ্ঠিত (held), জাতীয় (national), পুলিশ (police), উপজেলা (sub-district), হয়েছে (happened), আটক (arrested), নিহত (killed), সভাপতি (president)
COVID & Digital Services (করোনা ও ডিজিটাল পরিষেবা)	10.80%	বাংলাদেশ (Bangladesh), হয়েছে (happened), হালনাগাদ (updated), সাইটটি শেষ (site finished), জাতীয় (national), বাতায়ন (portal), করোনা ভাইরাস (coronavirus), প্রতিরোধে যোগাযোগ (contact for prevention), ডিজিটাল (digital), জলাভূমি উন্নয়ন (wetland development)
Technology & Online Services (প্রযুক্তি ও অনলাইন পরিষেবা)	7.20%	পারবেন (can), ভিডিও (video), টাকা (money), ডাউনলোড (download), আপনাকে (to you), প্রশ্ন (question), ক্লিক (click), পাবেন (will get), কিভাবে (how), নাম (name)
Daily Life & General Discussion (দৈনন্দিন জীবন ও সাধারণ আলোচনা)	6.80%	হয়ে (becomes), যায় (goes), একটা (one), দিয়ে (with), সময় (time), মানুষ (people), ভালো (good), মানুষের (of people), মা (mother), খাবার (food)
Finance & Banking (অর্থনীতি ও ব্যাংকিং)	9.20%	টাকা (money), ব্যাংক (bank), ব্যাংকের (of bank), প্রকল্প (project), লেনদেন (transaction), ঋণ (loan), মূল্য (value), শতাংশ (percentage)
News & Government Affairs (সংবাদ ও সরকারি বিষয়)	7.70%	হয়েছে (happened), দেশের (of country), খবর (news), গত (past), করোনা (corona), দেশে (in country), হয়েছে (has been), রয়েছে (exists), সরকার (government), বন্ধ (closed)

Figure 9. Topic modelling results for Bengali language in Chitrakshara-IL dataset.

Topic Name (English & Telugu)	Topic Ratio (%)	Top Words (with Translation)
Poetry & Literature (కవిత్వం & సాహిత్యం)	9.60%	వలె తారాడగ (like shining), లోపల ప్రాలేయచ్ఛాయల (inside icy shadows), కోరాడగ (like a wave), అలోలములాలోచనలేవేవో నా (deep thoughts in my mind), ఇన్ని (so many), అర్హత (qualification), deep, nature, అక్కడ (there), బంగారం ధర (gold price)
Personal Opinions & Social Media (వ్యక్తిగత అభిప్రాయాలు & సోషల్ మీడియా)	10.30%	నా (my), మా (ours), ఈ (this), వీరిచే పోస్ట్ (posted by them), ఏదో (something), పూర్తి ప్రొఫైల్‌ను (complete profile), నా చిన్ని ప్రపంచం (my little world), నా చిన్ని ప్రపంచానికి (to my little world), చెయ్యబడింది రాజ్యలక్ష్మి (done by Rajyalakshmi), లో (in)
Technology & Gadgets (సాంకేతికత & గాడ్జెట్‌లు)	7.80%	యువకుడు (young man), డి (D), డిస్ప్లే (display), అరుణాచలం యాత్రా (Arunachalam Yatra), కత్తితో (with knife), ప్రాసెసర్ (processor), విష్ణుకంచి (Vishnukanchi), ర్యామ్ (RAM), స్తోత్రాలు (hymns), కాణిపాకం (Kanipakam)
General Conversations & Thoughts (సాధారణ చర్చలు & ఆలోచనలు)	9.10%	ఈ (this), ఆ (that), అని (said), మీ (your), లో (in), చాలా (very), ఇది (this), నా (mine), కోసం (for), నుండి (from)
Movie Reviews & Comments (సినిమా సమీక్షలు & వ్యాఖ్యలు)	8.70%	తెర సినిమా (screen cinema), suresh_comments, ముఖ్యాంశాలు (highlights), sudheer_comments, comments, January, desk_comments, July, August, October
Entertainment & Films (వినోదం & సినిమాలు)	9.40%	ఈ (this), సినిమా (movie), లో (in), ఆ (that), తన (his/her), ఓ (a), చిత్రం (film), తో (with), చేసిన (did), ఇక (next)
News & Government Updates (వార్తలు & ప్రభుత్వ సమాచారం)	10.10%	ఈ (this), నుండి (from), కరోనా (Corona), ప్రభుత్వం (government), చేశారు (did), తెలంగాణ (Telangana), ఆయన (he), చేసిన (done), మంది (people), కోసం (for)
Online Services & Verification (ఆన్‌లైన్ సేవలు & ధృవీకరణ)	8.30%	మీ అభిప్రాయాలు (your opinions), తెలియజేసినందుకు ధన్యవాదాలు (thanks for sharing), దయచేసి (please), క్లిక్ చేయండి (click here), మీకు పంపించాము (sent to you), ధృవీకరణ కోసం (for verification), ఆ లింకుపై (on that link), ఈమెయిల్ ను (email), అన్ని (all), శ్రీ (Sri)
Politics & Business (రాజకీయాలు & వ్యాపారం)	11.20%	days, movies, బడ్జెట్ (budget), politics, దక్షిణ (South), వార్తలు (news), hrs (hours), క్రింద (below), •, అతిపెద్ద (biggest)
Sports & National News (క్రీడలు & జాతీయ వార్తలు)	9.50%	భారత (India), news, జట్టు (team), రైతులు (farmers), భారత్ (India), నుండి (from), టీమిండియా (Team India), కోరారు (requested), తొలి (first), చదువు సుఖీభవ (education happiness)

Figure 10. Topic modelling results for Kannada language in Chitrakshara-IL dataset.

Topic Name (English & Kannada)	Topic Ratio (%)	Top Words (with Translation)
Personal Thoughts & Conversations (ವ್ಯಕ್ತಿಗತ ಚಿಂತನೆಗಳು & ಮಾತುಕತೆ)	9.20%	ನಮ್ಮ (our), ನಿಮ್ಮ (your), ನನ್ನ (my), ನಾನು (I), ನೀವು (you), ನಾವು (we), ಅಂತ (like that), ನನಗೆ (to me), ನೋಡಿ (see), ಹೇಗೆ (how)
Regional News & COVID-19 (ಪ್ರಾದೇಶಿಕ ಸುದ್ದಿ & ಕೋವಿಡ್-19)	8.90%	ಕನ್ನಡ (Kannada), ಮಂಗಳೂರು (Mangalore), ಕೋವಿಡ್ (COVID), ಸಾವು (death), ಕರ್ನಾಟಕ (Karnataka), ಜಿಲ್ಲಾ (district), ಪೊಲೀಸರು (police), ಉಡುಪಿ (Udupi), ರಾಜ್ಯ (state), ಮಾಹಿತಿ (information)
Online Services & User Feedback (ಆನ್‌ಲೈನ್ ಸೇವೆಗಳು & ಬಳಕೆದಾರ ಪ್ರತಿಕ್ರಿಯೆ)	9.50%	ಕ್ಲಿಕ್ (click), ನಿಂದನಾತ್ಮಕ (negative), ದಯವಿಟ್ಟು (please), ಕಾಣಿಸಿಕೊಂಡರೆ (if visible), ನಾವು ಫಿಲ್ಟರ್ (we filter), ನಿಮ್ಮ ಅನಿಸಿಕೆ (your opinion), ಮಾಡಿದರೆ (if done), ನಿಯಮಗಳನ್ನು ಉಲ್ಲಂಘನೆ (violating rules), ಅಳವಡಿಸಿದ್ದೇವೆ (implemented), ವ್ಯಕ್ತಪಡಿಸಿದ್ದಕ್ಕೆ ಧನ್ಯವಾದಗಳು (thanks for expressing)
Family & Daily Life (ಕುಟುಂಬ & ದಿನನಿತ್ಯದ ಜೀವನ)	8.70%	ನನ್ನ (my), ತಂದೆ (father), ರಂದು (on date), ದಿನಾಂಕ (date), ಆನಿ (Ani), ಪುಟ್ಟ (small), ಬಂದು (came), ಅಂತಾ (like that), ನಾಲ್ವರು (four people), ಚಿತ್ರಗಳು (pictures)
Music, Yoga & Spirituality (ಸಂಗೀತ, ಯೋಗ & ಆಧ್ಯಾತ್ಮಿಕತೆ)	8.40%	ಸಂಗೀತ (music), ಸ್ಪಷ್ಟನೆ (clarification), ರೈ (Rai), ಯೋಗ (yoga), ಕ್ಷೇತ್ರಗಳಲ್ಲಿ (fields), ನಿತ್ಯ (daily), ಪುನೀತ್ (Puneeth), ಗಮನ (attention), ಗೋವಿಂದ (Govinda), ಅಪ್ಪು (Appu)
Politics & Elections (ರಾಜಕೀಯ & ಚುನಾವಣೆ)	10.10%	ಬಿಜೆಪಿ (BJP), ಕಾಂಗ್ರೆಸ್ (Congress), ಸಿದ್ದರಾಮಯ್ಯ (Siddaramaiah), ಬೆಂಗಳೂರು (Bangalore), ಯಡಿಯೂರಪ್ಪ (Yediyurappa), ಸಿನಿಮಾ ಪ್ರದರ್ಶನ (film screening), ಬೆಂಗಳೂರಿನ ಚಿತ್ರಮಂದಿರದಲ್ಲಿ (in Bangalore theater), ಪಕ್ಷದ (party's), ಸಿಎಂ (CM), ಮಾಜಿ (former)
Movies & Entertainment (ಚಲನಚಿತ್ರಗಳು & ಮನರಂಜನೆ)	9.80%	ಸಿನಿಮಾ (movie), ನಟ (actor), ಚಿತ್ರದ (of the film), ಚಿತ್ರ (film), ನಟಿ (actress), ದರ್ಶನ್ (Darshan), ಚಿತ್ರದ (in the film), ಕನ್ನಡ (Kannada), ಗೌಡ (Gowda), ಅಭಿಮಾನಿಗಳು (fans)
Government & National Issues (ಸರ್ಕಾರ & ರಾಷ್ಟ್ರೀಯ ವಿಷಯಗಳು)	8.60%	ವಿರುದ್ಧ (against), ಕೇಂದ್ರ (central), ಸರ್ಕಾರ (government), ಭಾರತ (India), ಭಾರತದ (of India), ನಂತರ (after), ಮೋದಿ (Modi), ಹೇಳಿದ್ದಾರೆ (said), ಕಳೆದ (last), ಇಂದು (today)
Social Change & Awareness (ಸಮಾಜ ಪರಿವರ್ತನೆ & ಜಾಗೃತಿ)	9.30%	ಸುದ್ದಿಗಳನ್ನು ನಾವು (we report news), ಸಮಾಜದ ಉತ್ತಮ (betterment of society), ಪ್ರಯತ್ನವನ್ನು ಮಾಡುತ್ತಿದ್ದೇವೆ (we are making an effort), ನಿಮ್ಮ ಮುಂದಿಡುವ (placing before you), ನೀವು ಸ್ವೀಕರಿಸಿದಾಗ (when you receive), ಪ್ರೋತ್ಸಾಹಿಸಿ ಸ್ವೀಕರಿಸಿ (encourage & accept), ತಲುಪಿಸುವವರು ನಾವಾಗಬಾರದೇಕೆ (why shouldn't we be messengers), ಒಳ್ಳೆಯ ಸುದ್ದಿಗಳಿಗೆ (for good news), ಸಮಾಜ ತೆರೆದುಕೊಂಡಿದೆ (society has opened up), ನಾವು ಬೆಳೆಯಬಹುದು (we can grow)
Finance & Market Trends (ಆರ್ಥಿಕತೆ & ಮಾರುಕಟ್ಟೆ ಪ್ರವೃತ್ತಿಗಳು)	7.50%	ನಿಮ್ಮ (your), ನೀವು (you), Kannada, ಬೆಲೆ (price), read, ಅಧಿಕ (high), ಕೆಳಗಿನ (below), ಚಿತ್ರದುರ್ಗ (Chitradurga), ಮಾಡುತ್ತದೆ (does), ಗಾಗಿ (for)

Figure 11. Topic modelling results for Telugu language in Chitrakshara-IL dataset.

Amou Haji – दुनिया का सबसे गन्दा इंसान, जब नहाया तो हो गयी मौत

अगर आपको कोई कहें कि एक साल तक आपको नहाना नहीं है या पानी से दूर रहना है तो ये आपको एक बेहूदा मजाक जैसा ही लगेगा। चलिए आज हम आपको एक ऐसे शख्स Amou Haji के बारे में बताते है जिसे दुनिया का सबसे गन्दा इंसान होने का...

ईट से बनी हुई छोटी सी झोपड़ी में रहने वाले हाजी मरे हुए जानवरों का सड़ा हुआ और बासी मांस खाते थे और पानी पीने के लिए भी वह जंग लगे आयल केन का इस्तेमाल करता थे। Amou Haji अपनी शक्ल सूरत से अनजान नहीं ...

साल 2013 में Amou Haji पर एक डाक्यूमेंट्री फिल्म भी बनायीं गयी थी जिसका नाम था ‘The Strange Life of Amou Haji’, जिसमें इनकी जीवनी के बारे में बताया गया था। जब ग्रामीणों के एक समूह उन्हें स्नान कराने के प्रयास...

Figure 12. Example Interleaved document for Hindi

బాహుబలిని కట్టప్పను ఎందుకు చంపారో అనే విషయాన్ని తెలుసుకోవడానికి గత రెండేళ్లుగా ఎదురుచూస్తున్నాం. నేను బాహుబలి వీర అభిమానిని. సాయంత్రం ఫస్ట్ షో చూడటానికి సోమవారం ఉదయం ఫ్లయిట్ ...

బాహుబలి1 సంచలన విజయం సాధించింది. పార్ట్1లో బాహుబలిని ఎందుకు చంపారనే ప్రశ్న మమ్మల్ని వెంటాడుతున్నది. దాంతో బాహుబలి2 చూడాలనే ఆసక్తి పెరిగింది. ఇండియాలోని చాలా మంది స్నేహితులు...

ఈ సినిమా చూడటం కోసం హసన్ ఖాన్ అనే పారిశ్రామిక వేత్త తన కుమారుడు, కూతురుతో కలిసి ఢాకా నుంచి కోల్‌కత్తాకు వచ్చారు. బంగ్లాదేశీయులకు బాలీవుడ్ సినిమాలు అంటే చాలా ఇష్టం. సౌత్ ఇండియా సినిమాల...

Figure 13. Example Interleaved document for Telugu

ನಿನ್ನೆ ಮಾಡಿಟ್ಟ ಹಿಟ್ಟಿನಿಂದ ಚಪಾತಿ ಮಾಡಿ ತಿಂತೀರಾ...? ಆರೋಗ್ಯಕ್ಕೆ ಮಾರಕ
ರೊಟ್ಟಿ (Roti), ಚಪಾತಿ ಅಥವಾ ಫುಲ್ಕಾ ಭಾರತೀಯ ಆಹಾರದ (Indian Food) ಬಹಳ ಪ್ರಮುಖ ಭಾಗವಾಗಿದೆ. ದಿನದ ಆಹಾರವಾಗಿರಲಿ ಅಥವಾ ರಾತ್ರಿ ಊಟವಾಗಿರಲಿ (Dinner), ಚಪಾತಿಯನ್ನು ...

ಜನ ಸಮಯ ಉಳಿತಾಯಕ್ಕಾಗಿ ಒಂದು ಬಾರಿ ಹಿಟ್ಟು ತಯಾರಿಸಿ. ಒಂದೇ ಹಿಟ್ಟಿನಿಂದ ಎರಡು ಅಥವಾ ಮೂರು ದಿನಗಳವರೆಗೆ ರೊಟ್ಟಿಗಳನ್ನು ತಯಾರಿಸುತ್ತಾರೆ, ಆದರೆ ಹಿಟ್ಟಿನ ಚಪಾತಿ ತಿನ್ನುವುದರಿಂದ ದೇಹಕ್ಕೆ ಹಾನಿಯಾಗುತ್ತದೆ (Health problem) ಎಂದು ನಾವು...

ವಿಜ್ಞಾನಿಗಳ ಪ್ರಕಾರ, ಹಿಟ್ಟನ್ನು ತಕ್ಷಣವೇ ಬಳಸಬೇಕು, ಇಲ್ಲದಿದ್ದರೆ ಇದು ಆರೋಗ್ಯಕ್ಕೆ ತುಂಬಾ ಹಾನಿಕಾರಕವಾದ ರಾಸಾಯನಿಕ (Harmful chemical) ಬದಲಾವಣೆಗಳನ್ನು ಉಂಟುಮಾಡುತ್ತದೆ. ಆಯುರ್ವೇದದಲ್ಲಿ ಇದನ್ನು ಹಾನಿಕಾರಕ ಎಂದೂ ಕರೆಯಲಾಗುತ್ತದೆ.

Figure 14. Example Interleaved document for Kannada

வெறும் விரைவில் ஜெனீவா மோட்டார் ஷோவில் பரபரப்பான பரந்த ன் மேற்பரப்பில் ஏஎம்ஜி ஜிடி எஸ் பதிப்பு மான்சொரி உலக பிரீமியர் பிறகு, புகழ்பெற்ற வாகன மெருகேற்றும் நிபுணர்கள் மீண்டும் ஆழமான கிணறு வெட்டித் ஒரு இனிய மாடலாக உற்பத்தி ஒரு புதிய வளர்ச்சி வழங்குகிறீர்கள்.

இனம் வேகத்தில் மூலைகளிலும் எடுக்க வேண்டும் – தனியாக சக்திவாய்ந்த பின் வலதுசாரி வெறி ஒரு ஈர்க்கக்கூடிய சான்றாக உள்ளது – மற்றும் திறன்: மான்சொரி வடிவமைப்பாளர்கள் சிறப்பு கவனம் டுரிஸ்மோ பின்...

ஆனால் அது மட்டும் சிறப்பாக மெருகேற்றும் வீட்டில் மறுவேலை என்று தனித்தனி பகுதிகள் வடிவம் ஆகும். குறிப்பாக இந்த மாதிரி, மான்சொரி அதன் உரிமையாளர், பந்தர் மூலம் வெளிப்படுத்தினர்...

Figure 15. Example Interleaved document for Tamil

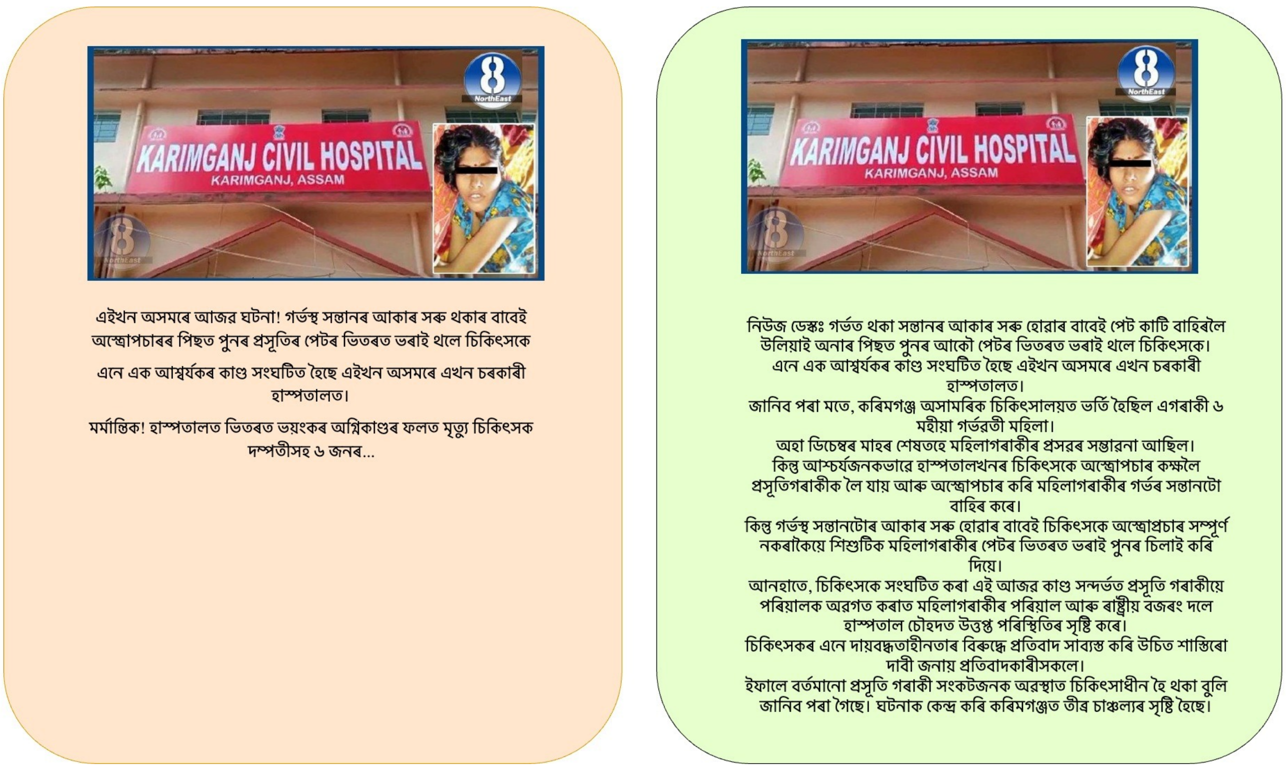

Figure 16. Comparison of the same interleaved document retrieved from mOSCAR against Chitrakshar-IL pipeline

Rank	Domain Name	# Docs
1	hi.medindia.net	346,395
2	hindi.news18.com	320,997
3	zeenews.india.com	307,467
4	hindi.alibaba.com	278,389
5	amarujala.com	273,655
6	thamilislam.blogspot.com	260,763
7	jmi.ac.in	249,841
8	dbhwd.portal.gov.bd	234,413
9	esakal.com	226,417
10	cpiup.blogspot.com	218,939
11	patrika.com	216,397
12	hindi.oneindia.com	213,376
13	newsnationtv.com	197,463
14	hindi.webdunia.com	177,184
15	hindutamil.in	175,280
16	vikatan.com	174,547
17	jagran.com	166,565
18	maalaimalar.com	162,471
19	khabar.ndtv.com	158,986
20	newstracklive.com	153,967
21	aajtak.intoday.in	147,650
22	myupchar.com	146,792
23	bhaskar.com	145,080
24	aajtak.in	132,935
25	dailythanthi.com	130,525
26	tamil.oneindia.com	127,532
27	livehindustan.com	127,211
28	raji-rajiworld.blogspot.com	125,848
29	malayalam.oneindia.com	119,674
30	kannada.oneindia.com	114,459
31	gujarati.oneindia.com	108,781
32	navbharattimes.indiatimes.com	108,541
33	pustak.org	106,866
34	telugu.oneindia.com	106,284
35	bengali.oneindia.com	104,611
36	anandabazar.com	103,539
37	lokmat.news18.com	103,088
38	udayavani.com	97,609
39	origin-www.amarujala.com	95,979
40	abpnews.abplive.in	94,350
41	tv9marathi.com	90,914
42	celebrity.astrosage.com	88,193
43	ndtv.in	86,399
44	loksatta.com	85,491
45	newstrack.com	84,154
46	vivalanka.com	84,075
47	prabhasakshi.com	83,619
48	hindi.newsbytesapp.com	82,588
49	mathrubhumi.com	79,742
50	m.jagran.com	78,261

Table 6. Top 1-50 URL domain names by number of documents in Chitrakshara-IL dataset.

Rank	Domain Name	# Docs
51	bhopalsamachar.com	77,566
52	tamil.news18.com	76,878
53	india.com	72,341
54	origin1qaz2wsx-hindi.webdunia.com	71,966
55	cgkhabar.com	69,293
56	mpbreakingnews.in	68,935
57	ap7am.com	68,798
58	lion-muthucomics.blogspot.com	68,121
59	mumbailive.com	67,404
60	hi.topwar.ru	64,629
61	newstm.in	62,938
62	hi.forvo.com	62,270
63	thejasnews.com	62,063
64	asianetnews.com	61,024
65	globaltamilnews.net	60,971
66	khaskhabar.com	60,394
67	naturalfoodworld.wordpress.com	57,818
68	hmtvlive.com	57,770
69	hindi.latestly.com	57,677
70	specialcoveragenews.in	56,047
71	ujjawalprabhat.com	55,647
72	pravakta.com	55,429
73	bn.fanpop.com	55,284
74	rokomari.com	55,252
75	matrubharti.com	54,751
76	tv9hindi.com	54,009
77	navodayatimes.in	54,009
78	pricedekho.com	53,973
79	sharechat.com	53,959
80	bharatkhabar.com	53,918
81	hindi.catchnews.com	53,878
82	ek-shaam-mere-naam.in	53,295
83	hindi.asianetnews.com	52,033
84	hindi.siasat.com	51,838
85	dinamalar.com	51,564
86	bsb.portal.gov.bd	51,367
87	merisaheli.com	50,811
88	varthabharati.in	50,529
89	upuklive.com	50,410
90	sarita.in	49,967
91	mymahanagar.com	49,775
92	swadeshnews.in	49,474
93	dw.com	48,642
94	thewirehindi.com	48,265
95	earchive.amarujala.com	47,870
96	marathi.webdunia.com	47,628
97	copypastelove.org	46,892
98	bansalnews.com	46,508
99	maayboli.com	45,694
100	liveaaryaavart.com	45,218

Table 7. Top 51-100 URL domain names by number of documents in Chitrakshara-IL dataset.