

Wasm: A Pipeline for Constructing Structured Arabic Interleaved Multimodal Corpora

Khalil Hennara
Misraj AI
Khobar, Saudi Arabia
hennara@misraj.ai

Ahmad Bastati
Misraj AI
Khobar, Saudi Arabia
bastati@misraj.ai

Muhammad Hreden
Misraj AI
Khobar, Saudi Arabia
hreden@misraj.ai

Mohamed Motasim Hamed
Misraj AI
Khobar, Saudi Arabia
hamed@misraj.ai

Zeina Aldallal
Misraj AI
Khobar, Saudi Arabia
aldallal@misraj.ai

Sara Chrouf
Misraj AI
Khobar, Saudi Arabia
chrouf@misraj.ai

Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
safwan@misraj.ai

Abstract

The performance of large language models (LLMs) and large multimodal models (LMMs) depends heavily on the quality and scale of their pre-training datasets. Recent research shows that LMMs trained on natural documents where images and text are interleaved outperform those trained only on image-text pairs across a wide range of benchmarks, leveraging advanced pre-trained models to enforce semantic alignment, image-sequence consistency, and textual coherence. For Arabic, however, the lack of high-quality multimodal datasets that preserve document structure has limited progress. In this paper, we present our pipeline *Wasm*[†] for processing the Common Crawl^{*} dataset to create a new Arabic multimodal dataset that uniquely provides markdown output. Unlike existing Arabic corpora that focus solely on text extraction, our approach preserves the structural integrity of web content while maintaining flexibility for both text-only and multimodal pre-training scenarios. We provide a comprehensive comparative analysis of our data processing pipeline against those used for major existing datasets, highlighting the convergences in filtering strategies and justifying our specific design choices. To support future research, we publicly release a representative dataset dump along with the full pipeline implementation.

1 Introduction

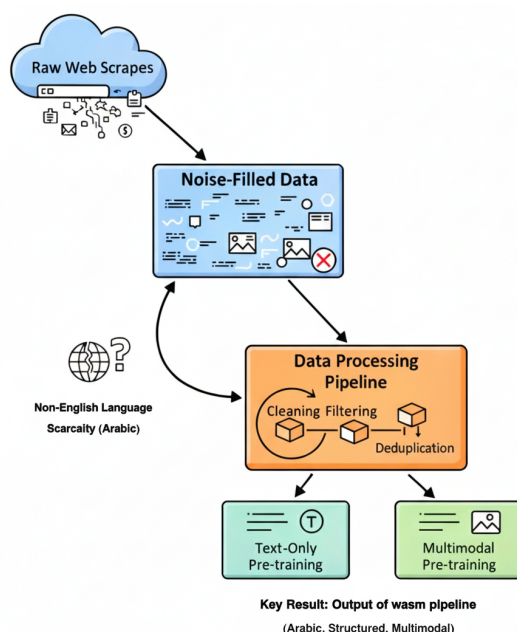


Figure 1: High-level overview of the *Wasm* pipeline, from Common Crawl WARC extraction to the final structured, Markdown-formatted output.

Arabic presents a compelling case study in low-resource Natural Language Processing (NLP). Despite its status as a major global language, high-

[†]*Wasm*: meaning 'tag' or 'mark,' which reflects the unique preservation of web markup structures in our data set.

^{*}<https://commoncrawl.org>

quality structured web corpora, particularly multimodal datasets, remain scarce. While data processing pipelines for English have matured considerably, Arabic-specific infrastructure and curation efforts have lagged behind, creating a significant bottleneck for regional AI development.

Recent studies emphasize that interleaved image–text data, which preserves the natural sequence of textual and visual elements within documents, is essential for training advanced multimodal models (Zhu et al., 2023; Laurençon et al., 2023; Li et al., 2024; Awadalla et al., 2024; Chen et al., 2025). Unlike isolated caption–image pairs, interleaved corpora capture document-level structure, allowing models to learn long-range dependencies, maintain narrative coherence, and align images with text across multiple segments or pages. This structure enables richer multimodal reasoning, such as following temporal or story-driven progressions, comparing various images within a context, and establishing complex instructions in visual evidence.

In contrast, existing Arabic resources (Aloui et al., 2024; Farhat et al., 2024; Nguyen et al., 2023) emphasize plain text extraction, discarding valuable structural cues such as document layout, formatting, and image associations, precisely the elements that interleaved corpora preserve and leverage for training advanced multimodal models. A recent effort, AMCrawl (Aboukazzana et al., 2025), addresses this gap by constructing a large-scale Arabic interleaved image-text corpus together with image-text pairs; however, it flattens documents into linear image-text sequences and does not preserve the underlying document structure (e.g., headings, lists, tables) or provide a structured, Markdown-based representation of the source pages.

In this paper, we present **Wasm**, an Arabic pipeline that produces both interleaved multimodal datasets and text-only corpora while fully preserving document structure Figure 1. To our knowledge, *Wasm* is the first Arabic pipeline to jointly preserve document-level structure (headings, lists, tables) and represent it as Markdown, offering a single output format that can be flexibly degraded into text-only or interleaved multimodal training data. *Wasm* retains document-level structure and the natural interleaving of text and images as they appear on their respective webpage, preserving layout cues and image–text associations needed for multimodal training. Figure 2 illustrates the struc-

tured output produced by our pipeline. Our work builds upon the OBELICS framework (Laurençon et al., 2023), adapting and extending it for Arabic to form an optimized pre-processing pipeline tailored to Arabic web data and multimodal use cases. Through systematic analysis of filtering techniques, corpus origins, and computational requirements across these datasets, our objective was to establish best practices for Arabic corpus construction and identify the most effective pre-processing strategies for multilingual and multimodal applications.

Our primary contributions are as follows.

1. We introduce *Wasm*, a new framework that processes and produces Arabic datasets. To our knowledge, it is the only Arabic pipeline that preserves structural information as Markdown while maintaining flexibility between text-only and interleaved multimodal outputs.
2. We release a part of the *Wasm* pipeline as open source¹, providing the Arabic-adapted module for interleaved multimodal unformatted data extraction. This component enables reproducible construction of Arabic web corpora with preserved text–image sequences and can serve as a foundation for further dataset development.
3. We provide a comprehensive comparative analysis of our data processing pipeline against those used for major existing corpora, highlighting the convergences in filtering strategies and justifying our specific design choices.
4. We release a Misraj Structured Data Dump (MSDD)² processed by *Wasm*.

2 Related Work

The creation of large-scale, high-quality datasets has been fundamental to recent breakthroughs in LLMs and LMMs. Dataset curation methodologies have evolved from simple crawling approaches to sophisticated multi-stage pipelines that balance scale with quality. This section reviews key developments in the construction of text-only and multimodal datasets, with particular attention to Arabic resources.

¹<https://github.com/misraj-ai/wasm>

²<https://huggingface.co/datasets/Misraj/msdd>

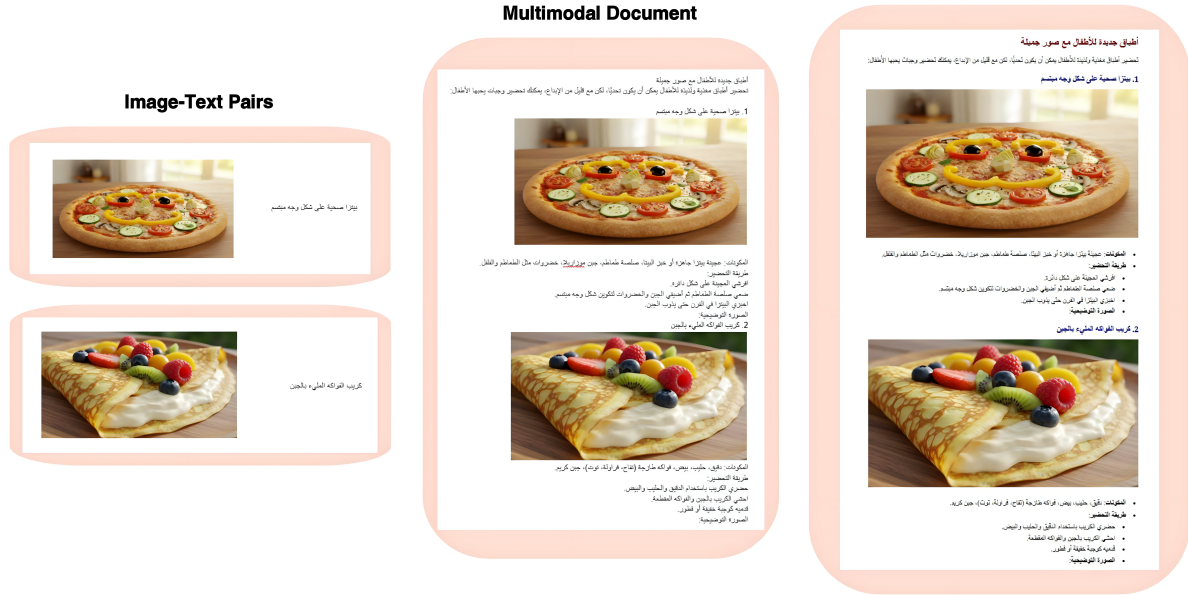


Figure 2: Example of the structured Markdown output produced by *Wasm* for a single web page, showing preserved headings, paragraphs, lists, and interleaved image placement.

2.1 Text-Only Corpora

Early large-scale text datasets established foundational principles for web data curation that remain influential today. The ROOTS corpus (Laurençon et al., 2022), developed for training BLOOM (Le Scao et al., 2023), exemplifies the construction of multilingual datasets by combining diverse sources such as Oscar (Ortiz Suárez et al., 2019), Common crawls, GitHub³ and more, with extensive source-specific filtering and deduplication.

Subsequent work has focused on scaling multilingual coverage while refining quality metrics. For example, CulturaX (Nguyen et al., 2023) aggregates the dumps of mC4 (Xue et al., 2020) and mOSCAR (Futrel et al., 2024) while incorporating perplexity-based scoring, repetition ratios, and confidence thresholds for language identification. These techniques have since become standard in large-scale corpus construction.

FineWeb (Penedo et al., 2024) represents a recent state-of-the-art corpus. Building on the Massive Text pipeline, it combines neural quality classifiers with custom heuristics and introduces per-dump MinHash deduplication, shown to outperform global deduplication in preserving quality at scale. FineWeb2 (Penedo et al., 2025) generalizes

this pipeline to more than 1000 languages, introducing a scalable multilingual dataset construction framework with language-adaptive filtering and deduplication. In addition, it proposes a principled rebalancing strategy based on duplication and quality metrics, yielding improved downstream model performance across diverse languages.

Arabic-specific curation efforts highlight how pre-processing pipelines adapt to the challenges of non-English web data. The 101 Billion Arabic Words corpus (Aloui et al., 2024) applied URL filtering, normalization (e.g., Unicode standardization), and document-level deduplication to Common Crawl WET⁴ files, while ArabicWeb24 (Farhat et al., 2024) incorporated more advanced methods such as Gopher quality filtering (Rae et al., 2021) and MinHash-based deduplication. However, like all existing Arabic text-only resources, they remain restricted to plain-text extraction, discarding structural and multimodal information present in native web documents.

2.2 Multimodal Corpora

The evolution toward Vision Language Models (VLMs) has necessitated datasets that preserve the rich contextual relationships between textual and visual content as they naturally occur on the web.

⁴These files are a subset of the WARC (Web ARChive) format, specifically holding "extracted plaintext".

³<https://github.com/>

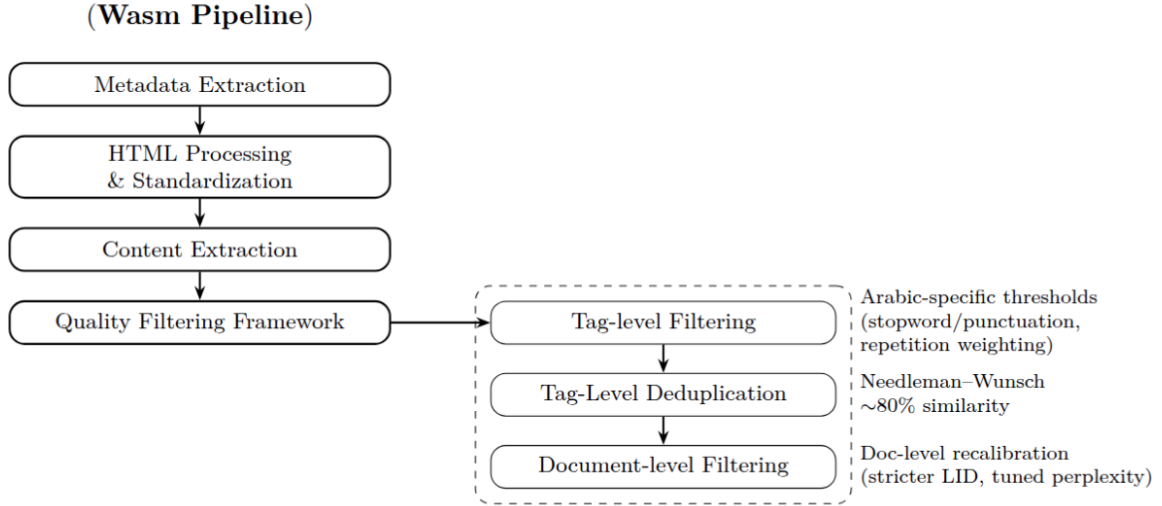


Figure 3: Detailed view of the *Wasm* pipeline stages: metadata extraction, HTML processing and standardization, content structuring into Markdown, and multilevel (tag- and document-level) quality filtering.

This requirement has driven methodological innovations beyond the simple extraction of image-text pairs. Early large-scale multimodal datasets, such as LAION-400M (Schuhmann et al., 2021) and LAION-5B (Schuhmann et al., 2022), were mainly based on large-scale capture of image capture pairs from Common Crawl, filtered using CLIP model (Radford et al., 2021). Although these resources proved invaluable for scaling multimodal training, their pair-based design eliminates broader document context and structural information.

In contrast, more recent work has focused on interleaved multimodal datasets that retain the sequential and structural interplay of text and images within documents. MMC4 (Zhu et al., 2023) advanced beyond basic co-occurrence by incorporating image-text alignment scores and document-level quality metrics into filtering pipelines. This work established the importance of semantic coherence in the construction of multimodal datasets. OBELICS (Laurençon et al., 2023) marked a further paradigm shift by prioritizing structural preservation: Rather than isolating image-text pairs, OBELICS maintains the interleaved document structure found in web content, creating a corpus of 141 million documents where images and text retain their natural sequential relationships. The processing pipeline operates at both document and HTML node levels, with final output representing documents as coherent sequences of text tokens and contextually-positioned images. Methodologically, OBELICS and MMC4 represent complementary approaches to HTML processing: OBELICS

leverages the Document Object Model (DOM) tree structure for comprehensive content filtering, while MMC4 uses HTML primarily for image location and integration with existing text corpora like C4, highlighting the trade-off between structural fidelity and processing efficiency.

More recently, OmniCorpus (Li et al., 2024) has pushed this paradigm further by introducing a 10-billion-level interleaved dataset, incorporating more diverse sources, including English and non-English web domains, as well as video-centric sites, and offers flexible formatting that can be degraded into pure text corpora or image-text pairs.

MNiT-1T (Awadalla et al., 2024) extends this trajectory by introducing a trillion-token multilingual multimodal corpus constructed from Common Crawl, including new source types such as PDFs and arXiv papers. It combines large-scale image-text pairing with document-level filtering and quality heuristics for robust VLM pre-training.

Complementing these scale-focused efforts, CoMM (Chen et al., 2025) addresses the qualitative limitations of existing interleaved datasets. It introduces a coherent interleaved multimodal corpus, applying multiperspective filters (text coherence, image sequence consistency, image-text alignment) that improve the quality of interleaved training data and enhance the in-context learning capabilities of multimodal LLMs.

For Arabic specifically, early multimodal resources were mainly based on pairwise, often translated, image-text datasets (Alwajih et al., 2024). More recently, AMCrawl (Aboukazzana et al.,

2025) introduced a web-scale Arabic interleaved image-text corpus of 2.8 million documents together with 5 million QA pairs, and used it to train a native Arabic vision-language model. AMCrawl demonstrates the feasibility of large-scale Arabic interleaved corpus construction, but, like MMC4, OBELICS, OmniCorpus, MINT-1T, and CoMM, it represents documents as flattened sequences of text and images rather than preserving the original structural hierarchy (headings, lists, tables) of the source page. This gap motivates *Wasm*, which, to our knowledge, is the first Arabic framework to jointly preserve full document structure and provide a Markdown-based representation that supports both text-only and interleaved multimodal pre-training.

3 Wasm Pipeline

This section details the comprehensive data processing pipeline developed to create a structured Arabic dataset. Our pipeline closely follows the methodology established by OBELICS (Laurençon et al., 2023), adapted specifically for Arabic language processing requirements. Figure 3 provides an overview of the pipeline architecture. All steps in the pipeline were iteratively tuned and refined through careful observation of the outputs on large samples of the data, ensuring that filtering thresholds, normalization rules, and structural preservation techniques were well-suited to the characteristics of Arabic web content.

3.1 Metadata Extraction

The initial phase involves extracting metadata by filtering web pages containing Arabic language content (not necessarily exclusively Arabic) from selected Common Crawl dumps. Each dump was processed separately to facilitate efficient separation and deduplication. We perform a language-based filtering before loading the Web Archive (WARC) files. This early intervention not only ensures that irrelevant data is excluded at the source but also saves substantial computational resources, including time, memory, and storage. Each scraped article is then associated with a set of metadata that facilitates storage, retrieval, and analysis. This metadata includes the URL of the article, the storage location within the corresponding Common Crawl snapshot as WARC files, and the byte position indicating where the webpage begins within that file. In addition, it records the size of the web-

page content in bytes, the detected language(s) of the webpage, and the domain name of the source website.

3.2 HTML Processing and Standardization

Using the extracted metadata (filename, offset, and length), the corresponding WARC files were accessed to retrieve the raw webpage content, which was subsequently converted to HTML format. To reduce noise and improve data quality, several pre-processing filters were applied. First, repeated line breaks and whitespace were normalized into single instances to ensure consistency in text formatting. The HTML comments, which do not have semantic value, were then eliminated. Finally, all CSS-related content was removed to eliminate styling artifacts that do not contribute to the linguistic or semantic properties of the data.

3.3 Simplifying and Structuring Web Content

Following HTML simplification, the content was converted into *Markdown* format to facilitate downstream text processing, preserve the document’s basic structural hierarchy, and allow precise extraction of textual and visual elements. Subsequently, the extraction process categorized the webpage content into two primary types: textual and visual. Our framework transforms texts into structured text; this step encompasses elements such as headers, paragraphs, ordered and unordered lists, and tables. The visual category comprises figures (*fig*) and image (*img*) tags. To maintain semantic coherence, text elements sharing identical tags are concatenated, thereby preserving both structural integrity and the contextual flow of the content.

3.4 Quality Filtering Steps

The pipeline employs a multilevel filtering system at both the tag and document levels to ensure data quality and relevance. A *tag* refers to any connected textual unit (e.g., paragraph, list, or section), while a *document* corresponds to the entire web page.

3.4.1 Arabic Perplexity Model

To provide a robust filtering strategy similar to the one introduced in (Laurençon et al., 2023), we trained a custom perplexity model specialized for Arabic text based on the KenLM (Heafield, 2011), which we called *WasmKenLM*. The training corpus for this 5-gram model comprises 37,258,520 samples across three domains: news, scientific, and

dialectal Arabic. This totals approximately 2.98 billion tokens, as measured by the Kuwain tokenizer (Hennara et al., 2025).

Table 1 presents the token distribution across the corpus categories.

Domain	Tokens
News	2,129,569,014
Scientific	515,017,029
Dialectal Arabic	337,945,783
Total	2,982,531,826

Table 1: Token distribution of the perplexity model training corpus by domain.

The KenLM model was trained with a 5-gram order using the `lmplz` toolkit and subsequently compiled into a binary format via `buildbinary` for efficient inference.

3.4.2 Tag-level Filtering

Compared to the baseline pipeline OBELICS, we introduced several manual modifications to adapt the filtering process to the specific characteristics of the Arabic language. These adjustments included relaxing or removing certain thresholds that were originally optimized for English but do not generalize well to Arabic.

For example, we reduce the weight of the *Word Repetition Ratio*, which measures the proportion of repeated words in a text. In Arabic, word repetition often carries stylistic and rhetorical significance rather than being indicative of low-quality content. Similarly, we removed the *Stopword Ratio* filter, as Arabic exhibits a rich vocabulary and flexible syntactic structures that allow grammatically correct sentences with relatively few function words. In the same spirit, the *Punctuation Ratio* was also discarded, since Arabic web content frequently lacks punctuation, and using this metric would disproportionately eliminate valid Arabic text. We also disabled the *Common Word Ratio* filter, which penalizes texts containing many high-frequency words. In Arabic, this would be biased against authentic content, given that the distribution of common words differs significantly from English. Likewise, the *Special Character Ratio* (e.g., emojis, abbreviations) was adjusted more lightly, since contemporary Arabic text often incorporates such elements without necessarily being low-quality. We applied a stricter *Language Identification* process to guarantee that the text is predominantly Arabic, while

still allowing for the natural occurrence of foreign terms (for example, English terminology).

3.4.3 Visual Data Filtering

Our image data filtering strategy was tailored to the characteristics of Arabic web content, with an emphasis on maximizing data retention while upholding quality standards. Given the relative scarcity of Arabic multimodal resources, we adopted a conservative approach that avoids unnecessary exclusions.

Instead of downloading images directly, we collected their URLs to reduce storage costs, accelerate the acquisition process, and enable scalable filtering. This design naturally shifted the focus of filtering from individual images to the URL level. To ensure safety and appropriateness, we maintain a blacklist of websites that host explicit or unsuitable material and exclude all associated image URLs. In particular, Arabic web content is rarely hosted on mainstream platforms that contain prohibited content, which further justifies this site-level filtering strategy.

The resulting set of URLs forms a flexible foundation for subsequent stages of image processing and task-specific filtering, allowing later adjustments to be aligned with the requirements of different models and training objectives.

3.4.4 Tag-Level Deduplication

We have worked to remove implicit duplicates within documents. For example, some sites contain duplicate ads. To address substantial repetition at the tag level observed in the dataset, we implemented the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) with a similarity threshold of 80% to efficiently identify and remove nearly duplicate content.

3.4.5 Document-level Filtering

At the document level, we applied the same set of filtering criteria as in the tag-level filtering 3.4.2, but with different parameter values. These values were recalibrated to reflect document-wide characteristics rather than paragraph-level ones. In particular, thresholds were tuned to balance the need for higher-quality long-form content with the goal of retaining as much Arabic data as possible.

3.5 Dataset Statistics

Table 3 summarizes the overall scale of the pipeline. Starting from approximately 2.5 million Arabic web documents identified in Common Crawl, the

Stage	Documents	Images
Raw extraction	2.5M (100%)	30.3M (100%)
Filtering & deduplication	1.76M (70.3%)	21.8M (71.9%)

Table 3: Document and image counts before and after the *Wasm* filtering and deduplication stages.

Dataset	Arabic	Interleaved	Struct. preserved	Dedup. granularity
MMC4 (Zhu et al., 2023)	✗	✓	✗	Document
OBELICS (Laurençon et al., 2023)	✗	✓	✗	Document (MinHash)
OmniCorpus (Li et al., 2024)	✗	✓	✗	Document
MINT-1T (Awadalla et al., 2024)	✗	✓	✗	Document
101 Billion Arabic (Aloui et al., 2024)	✓	✗	✗	Document
ArabicWeb24 (Farhat et al., 2024)	✓	✗	✗	Document (MinHash)
AMCrawl (Aboukozzana et al., 2025)	✓	✓	✗	Document
Wasm (ours)	✓	✓	✓	Tag / node-level

Table 4: Comparison of *Wasm* with existing Arabic and interleaved multimodal corpora. “Struct. preserved” refers to retention of document-level structural elements such as headings, lists, and tables, rather than flattening documents into linear text-image sequences.

authenticity of human-authored text.

For a rigorous assessment, we constructed an evaluation suite comprising four distinct datasets: FineWeb-2, Ara24, CulturaX, and Dataset 101. We randomly sampled 100 examples from each dataset, yielding 400 evaluation samples in total. Human annotators then reviewed whether each example was correctly classified by each model.

Filtering philosophy: high precision over high recall. Because the goal of the perplexity filter is to *remove* bad text without discarding good text, we frame the human evaluation as a binary detection task. For each example flagged as low quality, the annotator judges whether the flag is correct (a *True Positive*, TP) or whether the example is in fact acceptable text that was incorrectly removed (a *False Positive*, FP). The failure mode we most want to avoid is the latter: good Arabic text being silently discarded.

Accordingly, our filtering strategy is designed to **minimize False Positives** at the cost of tolerating more False Negatives (borderline low-quality content that slips through the filter). This is achieved by setting a conservative, high perplexity threshold: *WasmKenLM* only flags an example as low quality when it is highly confident, prioritising **Precision** over Recall:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

A high-precision filter guarantees that when *Wasm* removes an example, that removal can be trusted. We consider this trade-off appropriate for a low-

resource language setting: for Arabic, which constitutes only $\sim 0.6\%$ of Common Crawl content, over-filtering is considerably more costly than under-filtering.

Human review of the 400 sampled examples confirmed that *WasmKenLM* consistently outperforms the WikiKenLM baseline. WikiKenLM, trained solely on Arabic Wikipedia, showed limited sensitivity to the range of quality issues present in Arabic web text, frequently failing to flag incoherent or machine-generated content while in other cases incorrectly penalising authentic dialectal or informal text. *WasmKenLM*, by contrast, benefited from its multi-domain training corpus (news, scientific, and dialectal Arabic) and produced notably fewer False Positives across all four evaluation datasets, confirming the suitability of its thresholds for Arabic web curation.

4.2.2 Dataset Assessment and Exclusion Rates

Using the 4,000 manually reviewed samples described in Section 4.2.1, we analyzed the practical impact of our filtering thresholds. For each dataset, we calculated the *exclusion rate* the proportion of examples rejected for exceeding acceptable perplexity scores.

Table 5 reports the exclusion rates across the four datasets. The human review conducted on these samples confirmed the high accuracy of our model in successfully targeting low-quality content. Specific examples of text excluded by these thresholds are presented in Table 2. Ultimately, applying a specialized filtering technique like *WasmKenLM* yields significantly higher-quality data, which di-

rectly and positively affects the performance of downstream applications relying on these datasets.

Dataset	Exclusion Rate (%)
FineWeb-2	1.766
Ara24	7.820
CulturaX	8.605
Dataset 101	19.757

Table 5: Exclusion rates across datasets based on perplexity thresholds.

4.3 Granular Node-Level Deduplication.

While existing approaches typically perform deduplication at the document level (e.g., ROOTS (Laurençon et al., 2022), 101 Billion Arabic Words (Aloui et al., 2024), AM-Crawl (Aboukozzana et al., 2025)) or apply MinHash deduplication globally or per-dump (e.g., FineWeb (Penedo et al., 2024)), our methodology implements deduplication at the HTML node level. This granular approach enables the preservation of documents that contain unique content alongside duplicated elements (such as navigation menus or boilerplate text), significantly improving content diversity while maintaining processing efficiency. Node-level deduplication is particularly valuable for web documents where substantial unique content may coexist with repeated structural elements, allowing for more nuanced quality preservation than binary document-level decisions.

5 Conclusion

This study introduces *Wasm*, an Arabic multimodal processing framework that processes Common Crawl data, designed to preserve the structural and semantic integrity of web documents. To our knowledge, *Wasm* is the first Arabic pipeline to jointly preserve document-level structure and offer a Markdown-based output that flexibly supports both text-only and interleaved multimodal pre-training. *Wasm* provides a flexible foundation for training both LMMs and LLMs by maintaining document-level coherence, cross-modal alignments, and hierarchical structures such as captions, sections, and contextual dependencies. By releasing both the dataset and pipeline code, *Wasm* not only democratizes access to advanced multimodal Arabic resources but also pushes the boundaries of Arabic NLP development, enabling reproducible re-

search and laying the groundwork for future large-scale corpus construction. *Wasm* demonstrates that even for a language as syntactically rich and web underrepresented as Arabic, constituting only 0.6% of Common Crawl content careful pipeline design can yield a high-quality, structurally faithful multimodal corpus, offering a replicable template for other low-resource languages.

As an important direction for future work, we plan to use the Misraj Structured Data Dump (MSDD) to pre-train or continually fine-tune Arabic LLMs and LMMs, and to evaluate the resulting models on standard Arabic and multimodal benchmarks. Such experiments will provide direct empirical evidence of the downstream utility of structure-preserving, Markdown-based pre-training data compared to corpora that flatten document structure into linear text-image sequences.

References

- Shahad Aboukozzana, Muhammad Kamran J Khan, and Ahmed Ali. 2025. *AMCrawl: An Arabic web-scale dataset of interleaved image-text documents and image-text pairs*. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 448–465, Suzhou, China. Association for Computational Linguistics.
- Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. *101 billion arabic words dataset*.
- Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. *Peacock: A family of Arabic multimodal large language models and benchmarks*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.
- Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. 2024. 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.
- Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. 2025. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8073–8082.
- May Farhat, Said Taghadouini, Oskar Hallström, and Sonja Hajri-Gabouj. 2024. *Arabicweb24: Creating a high quality arabic web-only pre-training dataset*.

- Matthieu Futral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. *moscar: A large-scale multilingual and multimodal document-level corpus*. *arXiv preprint arXiv:2406.08707*.
- 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.
- Khalil Hennara, Sara Chrouf, Mohamed Motaism Hamed, Zeina Aldallal, Omar Hadid, and Safwan AlModhayan. 2025. *Kuwait 1.5 b: An arabic slm via language injection*. *arXiv preprint arXiv:2504.15120*.
- 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, et al. 2022. *The bigscience roots corpus: A 1.6 tb composite multilingual dataset*. *Advances in Neural Information Processing Systems*, 35:31809–31826.
- 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. 2023. *Obelics: An open web-scale filtered dataset of interleaved image-text documents*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. *Bloom: A 176b-parameter open-access multilingual language model*.
- Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. 2024. *Omni-corpus: A unified multimodal corpus of 10 billion-level images interleaved with text*. *arXiv preprint arXiv:2406.08418*.
- Saul B. Needleman and Christian D. Wunsch. 1970. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. *Journal of Molecular Biology*, 48(3):443–453.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. *Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages*.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures*. *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. *The fineweb datasets: Decanting the web for the finest text data at scale*. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. *Fineweb2: One pipeline to scale them all—adapting pre-training data processing to every language*. *arXiv preprint arXiv:2506.20920*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. *Learning transferable visual models from natural language supervision*. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. *Scaling language models: Methods, analysis & insights from training gopher*. *arXiv preprint arXiv:2112.11446*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. *Laion-5b: An open large-scale dataset for training next generation image-text models*. *Advances in neural information processing systems*, 35:25278–25294.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. *Laion-400m: Open dataset of clip-filtered 400 million image-text pairs*. *arXiv preprint arXiv:2111.02114*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. *mt5: A massively multilingual pre-trained text-to-text transformer*. In *North American Chapter of the Association for Computational Linguistics*.
- Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. *Multimodal C4: An open, billion-scale corpus of images interleaved with text*. *arXiv preprint arXiv:2304.06939*.