DIA2: A Comprehensive and Diverse Diacritized Modern Standard Arabic Corpus for Large-Scale NLP Research

Fatima Dekmak[1], **Shady Elbassuoni**[1], **Khaled Bashir Shaban**[2]
Hazem Hajj[1], **Wassim El Hajj**[3], **Yasmine Abu Adla**[1], **Buthaina Alabrash**[1]
[1] American University of Beirut, Lebanon
[2] Qatar University, Doha, Qatar
[3] American University of Beirut – Mediterraneo, Pafos, Cyprus
fkd04@mail.aub.edu, se58@mail.aub.edu, khaled.shaban@qu.edu.qa,
hh63@mail.aub.edu, wassim@aubmed.ac.cy, yaa41@mail.aub.edu,
bga07@mail.aub.edu

Abstract

The development of Arabic natural language processing (NLP) applications and large language models (LLMs) faces substantial challenges, primarily due to the scarcity of high-quality native Arabic datasets. To address this critical gap, we present DIA2 (a Comprehensive and Diverse Diacritized Modern Standard Arabic Corpus), a novel dataset curated from 28 diverse, carefully selected Arabic sources. DIA2 emphasizes the use of original Arabic text and explicitly avoids machine-translated content. The corpus incorporates substantial amounts of text from books, news articles, and poetry, and employs extensive data preprocessing to support NLP research and LLM development. Our preprocessing pipeline includes rigorous text cleaning, URL- and document-level deduplication, and automatic diacritization, while preserving a gold diacritized subset derived from manually annotated sources. The resulting corpus comprises over 140 GB of high-quality text, containing more than 26 million unique words and 41.9 billion tokens.To evaluate the proposed pipeline, we conducted controlled continued pretraining experiments using Llama3.1-8B on both raw and processed subsets of DIA2. The model trained on processed data consistently outperformed its counterpart across multiple Arabic evaluation benchmarks. These results highlight the positive impact of systematic preprocessing and the utility of DIA2 in empowering native Arabic LLMs and downstream NLP tasks.

Keywords: Dataset, Modern Standard Arabic, Large Language Models, Natural Language Processing, Preprocessing

1. Introduction

Arabic is one of the most widely spoken and culturally significant languages in the world, spoken by over 422 million people across 22 countries in the Middle East and North Africa (Habash, 2010; Farghaly and Shaalan, 2009). Despite rapid progress in natural language processing (NLP) for many high-resource languages, Arabic research remains comparatively underdeveloped, largely due to a combination of *morphological richness* (Habash, 2010; Farghaly and Shaalan, 2009), *dialectal diversity, orthographic variation* (El-Haj et al., 2014; Ibrahim et al., 2019; Ali, 2018), and limited availability of*large-scale, high-quality linguistic resources* (Darwish, 2018; Zaghouani, 2014; Alshaikh et al., 2021). As a result, most recent advances in NLP and large language models (LLMs) continue to disproportionately benefit high-resource languages (Wang et al., 2018; Ruder, 2018), leaving Arabic under-represented and constrained by fragmented, inconsistently curated, and often inaccessible datasets (Darwish, 2018; Zaghouani, 2014; Alshaikh et al., 2021).

Recent efforts have begun to address this gap through the release of large-scale Arabic corpora. However, many existing resources rely heavily on web-crawled or Common Crawl data, which can introduce substantial noise, duplication, and domain imbalance (Aloui et al., 2024; Farhat et al., 2024; Kreutzer et al., 2022). Other datasets incorporate machine-translated text (Sengupta et al., 2023; Ali et al., 2023), potentially distorting Arabic linguistic structure through cross-lingual artifacts. Moreover, most large Arabic corpora omit diacritics (Zerrouki and Balla, 2017; Fadel et al., 2019; Farghaly and Shaalan, 2009), despite their critical role in preserving syntactic, morphological, and semantic distinctions in Arabic, a morphologically rich language where meaning is often ambiguous without vocalization. These limitations collectively restrict the effectiveness of Arabic datasets for training robust language models and downstream NLP systems.

To address these challenges, we introduce DIA2 (a Comprehensive and Diverse Diacritized Modern Standard Arabic Corpus), an open-source, large-scale Arabic dataset explicitly designed to support high-quality NLP research and LLM development. Unlike many existing corpora, DIA2 is curated exclusively from native Arabic sources, avoiding reliance on machine-translated data. The dataset draws from 28 carefully selected sources, including books, news articles, encyclopedic content, and poetry, to ensure domain diversity and liguistic authentic-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 115–130
May 11, 2026.

ity. DIA2 is further distinguished by rigorous and transparent preprocessing pipeline, incorporating extensive data cleaning, URL- and document-level deduplication, and automatic diacritization, while preserving a gold diacritized subset derived from manually annotated resources[1].

The resulting corpus comprises over 140 GB of high-quality diacritized Arabic text, containing more than 26 million unique words and 41.9 billion tokens, making DIA2 one of the largest publicly available diacritized resources for Modern Standard Arabic (MSA)[2]. In addition to the diacritized version, a parallel non-diacritized variant is also released[3], enabling flexible use across a wide range of NLP tasks. To assess the practical utility of DIA2, we conduct controlled continued pretraining experiments using LLaMA 3.1 (8B) (Grattafiori et al., 2024).

2. Related Datasets

Large-scale Arabic corpora have recently emerged to address the scarcity of extensive datasets for Arabic language modeling. For instance, the 101 Billion Arabic Words Dataset (Aloui et al., 2024) targets authentic Arabic content and applies a rigorous cleaning pipeline to reduce noise and duplication. Similarly, the 1.5 Billion Words Arabic Corpus (El-khair, 2016), which is primarily composed of news articles, has been widely adopted in Arabic NLP research, although its scale remains limited for training contemporary LLMs.

More recent efforts aim to expand both scale and domain coverage. The Large Arabic Corpus for LLMs (Ali et al., 2023) integrates books and news sources, but also includes lower-quality web-scraped and machine-translated content, raising concerns about noise, mistranslations, and linguistic inconsistencies. Similarly, the Jais Dataset (Sengupta et al., 2023) aggregates large volumes of Arabic text but relies partly on translated and web-crawled data, potentially introducing cross-lingual artifacts and domain imbalance.

Other initiatives focus on refining Arabic-centric resources. AraMUS (Alghamdi et al., 2023) expands monolingual Arabic data by aggregating and preprocessing existing corpora, emphasizing scale and model training efficiency. While effective for large-scale pretraining, such approaches inherit limitations from their source datasets, including inconsistent cleaning and limited linguistic control.

In parallel, several task-specific Arabic datasets have been released, including KIND (Yamani et al., 2024), ArQuAD (Obeidat et al., 2024), CIDAR (Alyafeai et al., 2024), and ArabicaQA (Abdallah et al., 2024). These resources provide high-quality annotations for targeted tasks such as classification, reading comprehension, and question answering, but are narrow in scope and not designed for large-scale language model pretraining.

In summary, existing Arabic datasets contribute valuable resources across different scales and tasks; however, many suffer from one or more limitations, including heavy reliance on web-crawled data, inclusion of machine-translated text, lack of systematic deduplication, omission of diacritics, or restricted domain coverage. These gaps underscore the need for large-scale, high-quality, natively sourced, and consistently curated Arabic corpora. DIA2 directly addresses these limitations by combining native Arabic sources, extensive preprocessing, large-scale diacritization, and public availability, positioning it as a complementary and foundational resource for advancing Arabic NLP research and language modeling.

3. Dataset Overview

DIA2 is a large-scale corpus focused on MSA, desiged to address the persistent lack of high-quality, systematically curated Arabic datasets required for reliable NLP research and language model training (Aloui et al., 2024). The dataset was curated from 28 publicly available native Arabic sources, ensuring broad domain coverage and linguistic authenticity, while explicitly avoiding any machine-translated content (Obeidat et al., 2024). The sources span a broad geographic range, covering the Gulf (Oman, Saudi Arabia), the Levant, Egypt, North Africa (Morocco), as well as multi-regional resources, alongside classical Arabic resources representing the pre-Islamic and early Islamic periods. As shown in Table 1, the data sources are distributed across general-purpose content (72.93%), news (26.23%), books (0.50%), and poetry (0.35%).

The general-purpose category comprises diverse text types suitable for large-scale language modeling, including *CulturaX* (Nguyen et al., 2023), *NADiA* (Al-Debsi et al., 2019), *SANAD* (Einea et al., 2019), and *Arwiki* (Contributors, 2024).

The news category focuses on contemporary topics and includes several Arabic news corpora such as *AraNews* (Nagoudi et al., 2020), *Arabic News Articles* (Altamimi and Alayba, 2023), the *Ultimate Arabic News Dataset* (Al-Dulaimi, 2022), and the *Saudi Newspapers Corpus* (Alhagri, 2015).

DIA2 also incorporates formal and literary Arabic, including text from the *Hindawi Bookset* (Elfilali, 2023) and the *King Saud University Corpus of*

[1] https://huggingface.co/datasets/DIA2-Arabic/DIA2-Tashkeela-Gold

[2] https://huggingface.co/datasets/DIA2-Arabic/DIA2-Dataset-Diacritized

[3] https://huggingface.co/datasets/DIA2-Arabic/DIA2-Dataset-Non-Diacritized

General Purpose		News	
Source	**Size (MB)**	**Source**	**Size (MB)**
CulturaX (Nguyen et al., 2023)	190,000	NADiA (Al-Debsi et al., 2019)	21,100
Opus Wiki (Wenzek et al., 2020)	29,355	AraNPCC (Al-Thubaity et al., 2022)	20,000
ArWiki (Contributors, 2024)	10,600	SANAD (Einea et al., 2019)	19,100
Tashkeela (Zerrouki and Balla, 2017)	1,220	Arabic News Articles (Altamimi and Alayba, 2023)	17,500
KALIMAT (El-Haj and Koulali, 2013)	511	OSIAN (Zeroual et al., 2019)	6,490
KSUCCA (Alrabiah, 2014)	441	Arabic-News (Saad, 2019)	4,130
Al-Watan (Abbas et al., 2011)	28	Arabic Billion Words (Saad and Ashour, 2019)	1,800
Opus100 (Zhang et al., 2020)	3.49	Arabic Fake News (Khalil et al., 2022)	1,640
Arabic Quotes (Boulahia, 2023)	1.12	Goud-sum (Issam and Mrini, 2022)	622
		AraNews (Nagoudi et al., 2020)	632
		Ultimate Arabic News (Al-Dulaimi, 2022)	543
		Arabic Classification (Mohamed, 2018)	374
		Saudi Newspapers Corpus (Alhagri, 2015)	103
Books		**Poems**	
Source	**Size (MB)**	**Source**	**Size (MB)**
Sanadset (Mghari et al., 2022)	1,330	Arabic Poems 2 (Yousef et al., 2019)	528
Hindawi Books (Elfilali, 2023)	472	Arabic Poems 1 (Rezk, 2023)	438
		Ashaar (ARBML Team, 2024)	252
		Classical Arabic Poetry (El Karef, 2022)	49

Table 1: Raw Dataset Composition by Category and Source

Classical Arabic (KSUCCA) (Alrabiah, 2014). In addition, the dataset covers poetry and expressive language, drawing from the *Arabic Poem Comprehensive Dataset (Parts 1 & 2)* (Rezk, 2023;Yousef et al., 2019), *Ashaar* (ARBML Team, 2024), and *Classical Arabic Poetry* (El Karef, 2022).

A strict and multiligual-aware preprocessing pipeline was applied to remove noisy, duplicated, and sensitive content while preserving Arabic linguistic structure. The resulting corpus consists primarily of MSA, with limited inclusion of Classical Arabic to support stylistic and historical linguistic diversity. To address the scarcity of diacritized Arabic resources, DIA2 provides automatically diacritized text generated using the CATT model (Alasmary et al., 2024). In parallel, DIA2 preserves a gold diacritized subset derived from the manually annotated Tashkeela corpus (Fadel et al., 2019). Deduplication is applied at both the URL and document levels to ensure dataset uniqueness.

DIA2 is designed to support a wide range of Arabic NLP applications, including language modeling, diacritics restoration, text-to-speech, named entity recognition, and MSA versus dialect classification. Its scale, domain diversity, consistent preprocessing, and high-quality diacritization make DIA2 a robust monolingual resource for training and evaluating Arabic language technologies.

4. Dataset Preprocessing

To construct a high-quality Arabic corpus suitable for training modern NLP systems, comprehensive preprocessing is essential to remove noisy, redundant, and low-quality content. Such noise may arise from HTML tags, embedded JavaScript code, boilerplate text, or linguistic anomalies such as excessive punctuation, non-Arabic scripts, and unintelligible character sequences. These artifacts can disrupt Arabic linguistic structure and negatively impact model learning, making their removal a critical step. Accordingly, the proposed preprocessing pipeline in Figure 1 applies a sequence of filtering, cleaning, and normalization stages, including URL-based filtering, deduplication, and metric-based quality control. To support reproducibility and facilitate future research, the full preprocessing pipeline is made publicly available[4].

4.1. Data Cleaning

Inspired by prior large-scale Arabic preprocessing efforts (Ghaddar et al., 2022; Sengupta et al., 2023), we implemented a fine-grained, sentence-level data cleaning pipeline aimed at systematically eliminating noise while preserving linguistically meaningful content.

4.1.1. HTML and JavaScript Removal

HTML markup and JavaScript code are common artificats in open-source datasets derived from web sources and can significantly interfere with downstream text processing. These elements were removed using regular expressions (RegEx) and the

[4] https://github.com/fatimadekmak/DIA2-arabic-preprocessing

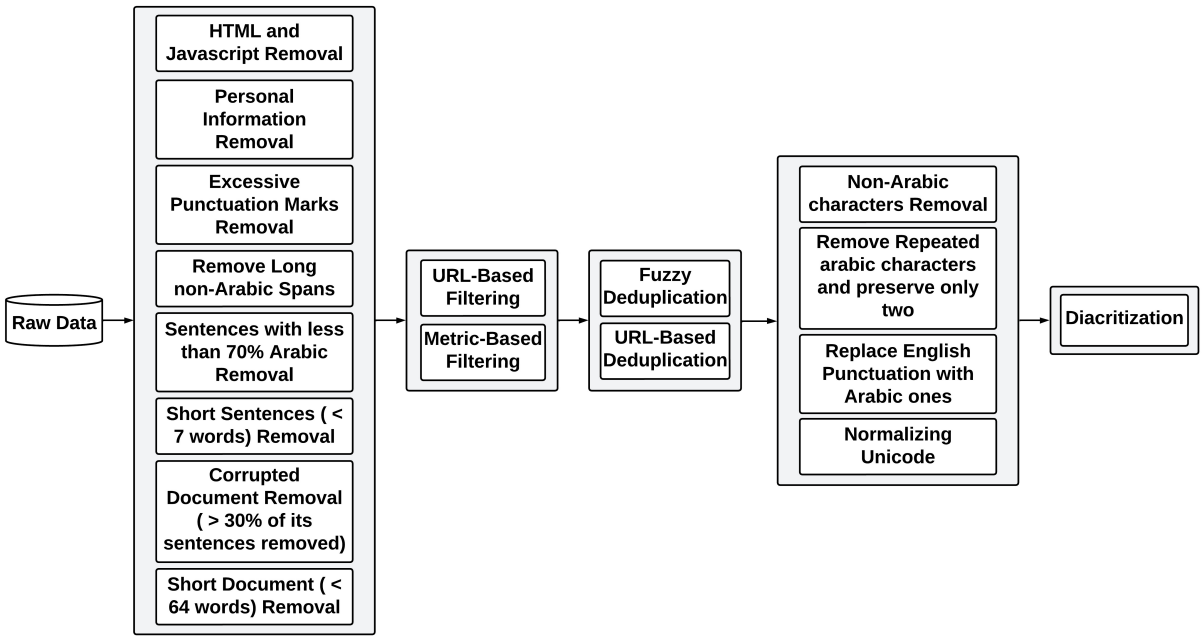

Figure 1: Data Preprocessing Pipeline

BeautifulSoup library[5]. RegEx patterns were utilized to identify and eliminate JavaScript segments (e.g., `<script>...</script>`) while BeautifulSoup was used to extract clean textual content from HTML documents, ensuring robust separation between text and markup.

4.1.2. Personal Information Removal

Open-source corpora may inadvertently contain personal information, raising privacy and ethical concerns. DIA2 mitigates this risk through targeted anonymization, replacing identifiable patterns such as phone numbers and email addresses with standardized placeholders (e.g., +999-999-9999 and Example@mail.com). Phone numbers are detected using RegEx patterns, while the phonenumbers Python library[6] was employed to improve detection robustness across formats.

4.1.3. Excessive Punctuation Marks Removal

Excessive punctuation can introduce artificial patterns and reduce textual coherence. DIA2 applies regular-expression-based filtering to remove punctuation sequences exceeding three consecutive characters, thereby improving readability while preserving legitimate expressive usage.

[5] https://pypi.org/project/beautifulsoup4/

[6] https://pypi.org/project/phonenumbers/

4.1.4. Enhancing Language Consistency

To maintain Arabic linguistic integrity, two complementary filtering strategies were applied. Regular expressions were used to identify and remove long non-Arabic spans, while selectively preserving essential elements such as numerals and commonly used symbols. This step also helped eliminate residual artifacts such as URLs, markup fragments, and code snippets that persisted after HTML and JavaScript removal. In addition, sentences containing less than 70% Arabic characters were removed to ensure that retained content is predominantly Arabic, effectively filtering out multilingual noise and irrelevant text.

4.1.5. Ensuring Data Quality

Additional quality constraints were applied to improve dataset coherence. Sentences shorter than eight words were removed due to limited informational value, and documents containing fewer than 64 words were discarded to ensure sufficient contexual content. Furthermore, documents in which more than 30% of sentences were removed during preprocessing were considered fragmented and excluded.

4.2. URL-Based Filtering

To ensure content appropriateness and ethical alignment, URL-based filtering was employed to exclude sources associated with unsafe content. We

utilized the UT15 blacklist[7], which has been widely adopted in large-scale dataset filtering (Nguyen et al., 2023; Penedo et al., 2023).

4.3. Metric-Based Filtering

Metric-based filtering was applied to identify and remove gibberish and low-quality text using a set of heuristic criteria. These included excessive character repetition, minimum Arabic word thresholds, validation against common Arabic vocabulary, and constraints on Arabic character proportions. Text containing non-Arabic scripts, such as Persian (Farsi), was further filtered using language identification tools (e.g., langdetect[8]). Collectively, these steps removed up to 35% of noisy entries, substantially improving dataset quality, consistent with prior large-scale preprocessing findings.

4.4. Arabic-Specific Cleaning Steps

Beyond general filtering, Arabic-specific normalization steps were applied to ensure linguistic consistency and suitability for Arabic NLP tasks.

4.4.1. Removing Non-Arabic Characters

An allowlist of valid characters was defined, including Arabic script, Arabic and Latin numerals, punctuation, and diacritics. All other characters were removed to maintain Arabic-centric textual integrity.

4.4.2. Removing Repeated Arabic Characters

Character repetition is common in Arabic, often used for emphasis or stylistic expression. While previous work removed all repeated characters (Hegazi et al., 2021; Mubarak and Darwish, 2014), our manual inspection revealed that many valid Arabic words naturally contain double character occurrences. To balance normalization and linguistic authenticity, DIA2 allows up to two consecutive occurrences of the same character, preventing excessive repetition while avoiding semantic distortion.

4.4.3. Punctuation and Unicode Normalization

Following (Sengupta et al., 2023), English punctuation marks were replaced with their Arabic counterparts to ensure textual consistency. For instance, the question mark (?) was replaced with (؟), the semicolon (;) with (؛), and the comma (,) with (،). In addition, Unicode normalization was applied to ensure consistent representation of Arabic characters, reducing encoding variability and simplifying downstream processing (Aloui et al., 2024).

4.5. Diacritization

While most Arabic corpora remove diacritics to simplify representation, this practice discards linguistically critical information. Given the objectives of DIA2, diacritics were explicitly restored to preserve syntactic and semantic distinctions.

A comparative evaluation was conducted to select an appropriate diacritization system, comparing CAMeL[9] and CATT (Character-based Arabic Tashkeel Transformer) (Alasmary et al., 2024). We adopted the encoder-only variant of CATT, which offers substantially faster inference with minimal performance degradation, making it suitable for large-scale processing.

Since CATT is a character-level model trained on clean MSA text, out-of-distribution inputs such as raw digits and non-MSA characters are incompatible. Accordingly, Persian-origin characters are normalized to their MSA equivalents and numeric tokens are expanded to Arabic word forms via num2words[10]. Original digits are restored post-diacritization through index mapping. Sequences exceeding 1,024 characters are segmented at punctuation boundaries.

4.5.1. Evaluation Setup

We sampled 150 sentences from DIA2 and restored diacritics using both tools, resulting in 300 sentences for evaluation. Three native Arabic annotators participated, each reviewing 200 sentences, with 100 sentences shared across annotators to enable inter-annotator agreement analysis.

Annotators were presented with a random mix of outputs and manually counted the number of incorrect diacritics, incorrect case-ending diacritics (الإعراب), and incorrectly diacritized words. Annotators also reported confidence for each judgment, which was later aggregated as an indicator of annotation reliability.

4.5.2. Inter-Annotator Agreement

Error counts were discretized into five bins: 0, 1, 2–4, 5–7, and 8+. This binning reflects error severity rather than precise counts and is suitable for comparative evaluation. Inter-annotator agreement was measured using Krippendorff's α (ordinal) (Krippendorff, 2011), which supports multiple annotators and does not assume equal distances

[7] https://dsi.ut-capitole.fr/blacklists/

[8] https://pypi.org/project/langdetect/

[9] https://camel-tools.readthedocs.io/en/latest/api/disambig/mle.html

[10] https://pypi.org/project/num2words/

between categories. As shown in Table 2, agreement scores ranged from 0.676 to 0.705 across the three evaluated dimensions, indicating acceptable agreement for exploratory and comparative analysis.

Annotator self-reported confidence rates were high overall (all above 92%), indicating that the task was well defined and the annotations are reliable for comparative evaluation.

Metric	Krippendorff's α	Interpretation
Incorrect Diacritics	0.705	Good
Incorrect Case Endings	0.680	Good
Incorrectly Diacritized Words	0.676	Good

Table 2: Inter-Annotator Agreement Scores (Krippendorff's α (Ordinal))

4.5.3. Tool Comparison Results

Using the human annotations, we computed four evaluation metrics: diacritic error rate, case-ending error rate, word error rate, and diacritization coverage. The results are shown in Table 3. Across all error-based metrics, CATT consistently produced fewer errors than CAMeL, while also achieving near-complete coverage of diacritized words. In contrast, CAMeL left a larger proportion of words undiacritized. Based on these results, CATT was selected as the diacritization tool for DIA2. All automatically diacritized versions of DIA2 released therefore rely on CATT Encoder-only outputs.

Metric	CAMeL	CATT
Diacritic error rate ↓	0.095	**0.036**
Case-ending error rate ↓	0.231	**0.081**
Word error rate ↓	0.269	**0.125**
Coverage rate ↑	0.936	**0.998**

Table 3: Diacritization Performance Comparison Between CAMeL and CATT

4.6. Deduplication

To prevent redundancy and reduce memorization effects during training, deduplication was applied at both URL and document levels. Duplicate entries originating from the same URL were removed by retaining a single instance per source. Since URLs were available for only a subset of the data, document-level deduplication was additionally applied. This step was performed using the datasketch[11] library, utilizing MinHash and Locality-Sensitive Hashing (LSH) for efficient near-duplicate detection. A sharding strategy inspired by (Öhman et al., 2023) enabled scalable intra- and inter-shard deduplication. The process was controlled by two parameters: the number of permutations for similarity estimation and a Jaccard similarity threshold for duplicate detection. The rationale behind these choices is discussed in Appendix B. Overall, this process reduced the dataset size by approximately 25%, from 70,522,230 to 52,449,241 rows.

[11] https://pypi.org/project/datasketch/

5. Dataset Evaluation

5.1. Dataset Size Reduction

The cumulative effect of the processing steps described in the previous section resulted in a substantially more compact and highter-quality, with the final version being 63.5% smaller than the original raw data. This reduction reflects the systematic removal of noisy, duplicated, and low-quality content, rather than indiscriminate filtering.

As illustrated in Figure 2, the dataset size decreased progressively across preprocessing stages. Starting from an initial size of approximately 329.26 GB, the data volume was reduced after each step, reaching a minimum of around 94.68 GB following deduplication. The subsequent diacritization stage increased the dataset size, due to the addition of vocalization marks, resulting in a final corpus size of approximately 140 GB.

In terms of instance counts, the dataset was reduced from 115,895,312 rows to 52,449,241 rows. Of the retained data, approximately 80% consists of long-form text (e.g., articles and book chapters), while the remaining 20% corresponds to shorter texts, including poems, quotations, and brief passages. The final corpus contains 18,376,574,554 words and 26,266,877 unique words. The average document length is 17 sentences, with an average sentence length of 26 words. Tokenization using the LLaMA 3.1-8B tokenizer (Grattafiori et al., 2024) yields a total of **41,931,425,578** tokens.

In addition to the automatically diacritized corpus, DIA2 includes a clean gold diacritized subset derived from the Tashkeela corpus (Fadel et al., 2019). This subset comprises **50,000 rows** (around **4.8 million tokens**) preserved with their original manual diacritics. Cleaning this subset resulted in an approximately 9% reduction in size, primarily due to the removal of short or noisy entries.

5.2. Comparison with Other Datasets

Table 4 presents a comparative overview of DIA2 against several widely used Arabic datasets, including 101 Billion Arabic Words (Aloui et al., 2024), the 1.5 Billion Words Arabic Corpus (El-khair, 2016), the Jais Dataset (Sengupta et al., 2023), AraMUS (Alghamdi et al., 2023), and ArabicWeb24 (Farhat

Feature	101BW	1.5BW	LDAC	Jais	Aramus	AW24	**DIA2**
Dialectal Content	✓	✓	✓		✓		
Diacritization							✓
Free of Translated Data	✓	✓			✓	✓	✓
Data Cleaning	✓		✓	✓	✓	✓	✓
Deduplication	✓			✓	✓	✓	✓
Code Availability						✓	✓
Public Availability	✓	✓				✓	✓
Includes Curated High-Quality Data			✓	✓	✓		✓

Table 4: Quantitative Comparison of DIA2 with other Arabic datasets: 101BW (101 Arabic Billion Words), 1.5BW (1.5 Billion Words Arabic Corpus), LDAC (Large and Diverse Arabic Corpus for LLMs), JAIS (JAIS Dataset), ARAMUS (Aramus Corpus), and AW24 (ArabicWeb24)

Feature	101BW	1.5BW	LDAC	Jais	Aramus	AW24	**DIA2**
Word Count	101B	1.5B	N/A	72B tokens*	N/A	26B tokens*	**18.4B**
Unique Words	N/A	3.3M	N/A	N/A	N/A	N/A	**26.3M**
Total Rows/Texts	89.1M	5.2M	N/A	N/A	N/A	86.8M	**52.5M**
Data Size	400GB	16GB	500GB	N/A	529GB	199 GB	**140GB**

Table 5: Quantitative Comparison of DIA2 with other Arabic Datasets
*: Token count was provided rather than the word count.

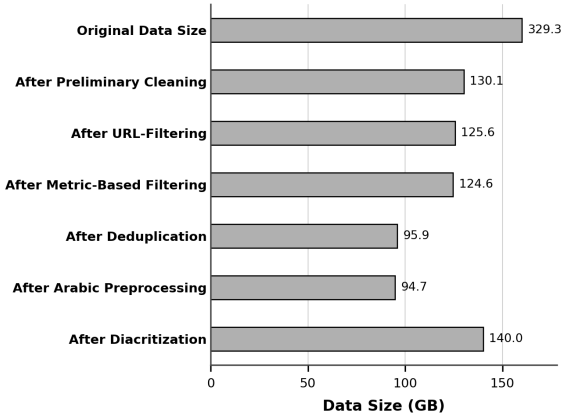

Figure 2: Change in Data Size of DIA2 across Processing Steps

et al., 2024). The comparison highlights key characteristics such as dialectal coverage, diacritization, reliance on translated data, data cleaning and deduplication, as well as code and data availability.

As shown in Table 4, most existing corpora lack diacritization, rely largely on web-crawled or machine-translated content, and offer limited preprocessing transparency or public availability. DIA2 addresses these gaps: it includes curated high-quality data, provides large-scale diacritization as a distinctive feature, applies rigorous data cleaning and multi-level deduplication, and releases both the dataset and preprocessing pipeline publicly. These characteristics collectively position DIA2 as a complementary, higher-fidelity resource for Arabic language modeling and evaluation.

6. Preprocessing Pipeline Evaluation

To assess the effectiveness of the proposed preprocessing pipeline, we conducted a controlled continued pretraining experiment in which the same base model was trained on raw and preprocessed subsets of DIA2 under identical conditions. The objective of this experiment is not to achieve state-of-the-art performance, but rather to isolate the impact of preprocessing on model behavior. Continued pretraining was performed on a limited subset of the dataset due to computational constraints, while maintaining sufficient scale for reliable relative comparison.

6.1. Pretraining Methodology

We performed continued pretraining of LLaMA 3.1-8B using Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning approach (Hu et al., 2021). Training was implemented using

Model	ArabicMMLU							Arabic Leaderboard		TruthfulQA Arabic	
	General	Grammar	High Arabic	Middle Arabic	Primary Arabic	Copa Arabic	AraDiCE PIQA MSA	Light	Complete	MC1	MC2
LLaMA-3.1-8B	0.577	0.515	0.344	0.519	0.433	0.629	0.592	0.453	0.458	0.287	0.454
Pretrained on DIA2											
DIA2 (raw)	0.525	0.479	0.328	0.333	0.345	0.618	0.588	0.452	0.454	0.274	0.444
DIA2 (prep.)	0.528	0.482	**0.349**	**0.630**	0.389	0.596	0.584	**0.457**	0.454	**0.310**	**0.476**

Table 6: Evaluation results of the original LLaMA-3.1-8B model and its continued pretraining variants on DIA2 (raw and preprocessed subsets) across multiple Arabic NLP benchmarks

Hugging Face Transformers, bitsandbytes for mixed-precision optimization, and accelerate for distributed execution, on a multi-GPU setup consisting of three NVIDIA GeForce RTX 4060 Ti GPUs (16 GB each).

Since the model is already multilingual and pretrained on diverse corpora, no tokenizer retraining or vocabulary expansion was performed. Continued pretraining was conducted exclusively on a subset of the non-diacritized version of DIA2, ensuring that performance differences stem from data quality rather than representation changes.

The training dataset comprised of 166K rows, with a batch size of 4 and gradient accumulation over 8 steps. The maximum sequence length was set to 512 tokens, and training was conducted for three epochs. LoRA was configured with a rank parameter of 8, using the 8-bit AdamW optimizer and a learning rate of 5×10^{-5}.

6.2. Performance Comparison

Table 6 reports the performance of the original multilingual LLaMA 3.1-8B model alongside its continued-pretraining variants trained on raw and preprocessed DIA2 data. Evaluation was conducted using the lm-eval-harness[12] framework across multiple Arabic benchmarks, including ArabicMMLU (Lakim et al., 2024), the Open Arabic LLM Leaderboard benchmarks (El Filali et al., 2024), and TruthfulQA Arabic (Lakim et al., 2024), covering general knowledge, grammatical competence, reasoning, and task-specific understanding.

Across the majority of benchmarks, the model pretrained on the preprocessed DIA2 subset consistently outperformed the variant trained on raw data. The largest relative improvements were observed on **High Arabic (+0.021)**, **Middle Arabic (+0.297)**, and **MC2 (+0.032)**, suggesting that cleaning and data normalization improve model robustness under resource-constrained continued pretraining.

Notably, performance gains were not uniform across all benchmarks. The preprocessed dataset did not yield improvements on **COPA Arabic** and **AraDiCE PIQA MSA**, both of which are translated from English. These benchmarks may introduce cross-lingual artifacts and alignment inconsistencies, limiting the benefits of Arabic-specific pretraining and reducing evaluation sensitivity to improvements in native Arabic data quality.

While the original multilingual model performed competitively overall, it showed weaker performance on several Arabic-specific benchmarks, particularly High Arabic and Primary Arabic. In contrast, the model trained on the preprocessed DIA2 subset achieved the highest scores on High Arabic (0.349), Middle Arabic (0.630), and the Arabic Leaderboard (light version) (0.457), indicating improved modeling of Arabic linguistic structure following exposure to cleaner, natively curated data.

Due to computational constraints and limited public availability of comparable datasets, we did not conduct continued pretraining experiments on alternative Arabic corpora listed in Table 4. Nevertheless, the consistent performance advantages observed when using preprocessed DIA2 data provide empirical evidence that rigorous preprocessing materially improves model behavior, even under limited training budgets.

7. Conclusion and Future Work

This work introduces DIA2, a large-scale, natively sourced, and diacritized corpus for MSA, designed to address longstanding limitations in Arabic NLP resources. DIA2 combines broad domain coverage, linguistic authenticity, systematic preprocessing, and large-scale diacritization, resulting in a dataset that supports both foundational language modeling and downstream Arabic NLP tasks. By explicitly avoiding machine-translated content and applying multi-level deduplication and Arabic-specific normalization, DIA2 offers a high-fidelity alternative to existing large-scale Arabic corpora.

We further demonstrated the value of DIA2 through a controlled preprocessing pipeline evaluation, showing that continued pretraining on preprocessed DIA2 subsets outperforms training on raw data across multiple Arabic benchmarks. These results empirically validate the importance of rigorous data curation and preprocessing for underresourced languages such as Arabic. Notably, im-

[12] https://github.com/EleutherAI/lm-evaluation-harness

provements were most pronounced on native Arabic benchmarks, underscoring the limitations of translated evaluation sets for accurately measuring Arabic language understanding.

Looking forward, several directions remain open for future work. First, expanding DIA2 to include additional high-quality dialectal Arabic data, while maintaining strict quality and provenance controls, could further enhance its applicability. Second, scaling continued pretraining to larger portions of the corpus and higher-capacity models would enable a more comprehensive assessment of DIA2's impact on foundation model performance. Third, systematic ablation studies of individual preprocessing components could provide deeper insight into their relative contributions. Finally, human-in-the-loop evaluation and error analysis, particularly for diacritization quality and semantic fidelity, would complement automated benchmarks and support more nuanced model assessment.

By releasing DIA2 and its preprocessing pipeline, we aim to lower barriers to high-quality Arabic NLP research and serve as a foundational resource for future advances in Arabic language modeling, evaluation, and responsible AI development.

8. Limitations

Despite the scale and rigor of DIA2, several limitations should be acknowledged. First, although the dataset draws from 28 diverse native Arabic sources, its content remains predominantly MSA. While this design choice ensures linguistic consistency and broad applicability, it limits direct coverage of regional Arabic dialects, which are increasingly relevant for real-world NLP applications.

Second, the majority of the corpus is automatically diacritized using the CATT model, and therefore may contain residual diacritization errors, particularly in out-of-context sentences, rare proper nouns, and domain-specific terminology. Although a gold diacritized subset is preserved for reference and evaluation, the scale of manually annotated diacritized data remains limited.

Third, the preprocessing pipeline relies on heuristic thresholds (e.g., minimum Arabic character ratios, sentence and document length constraints), which, while informed by prior work and empirical inspection, may inadvertently remove some valid but unconventional Arabic text, such as short-form expressions, quotations, or stylistically creative content.

Fourth, the evaluation of the preprocessing pipeline is based on continued pretraining under constrained computational resources, using a single base model (LLaMA-3.1-8B) and a limited subset of the dataset. While the results provide controlled evidence of preprocessing benefits, they do not fully capture the potential impact of DIA2 when used for full-scale pretraining or larger model architectures.

Finally, several Arabic benchmarks used for evaluation are translated from English, which may obscure improvements gained from native Arabic data curation and introduce cross-lingual artifacts that limit interpretability of results. This highlights the broader need for high-quality, natively constructed Arabic evaluation benchmarks.

These limitations point to clear and actionable directions for future work, many of which are already outlined in the preceding section, and do not detract from DIA2's primary contribution as a large-scale, transparently curated, and publicly available Arabic corpus.

9. Bibliographical References

M. Abbas, K. Smaili, and D. Berkani. 2011. Evaluation of topic identification methods on arabic corpora. *Journal of Digital Information Management*, 9(5):185–192.

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. arXiv preprint arXiv:2403.17848.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, and Alia O. Bahanshal. 2022. AraNPCC: The Arabic newspaper COVID-19 corpus. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 32–40, Marseille, France. European Language Resources Association.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. Catt: Character-based arabic tashkeel transformer. arXiv preprint arXiv:2407.03236.

Asaad Alghamdi, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang, Yi Zheng, Mehdi Rezagholizadeh, and Baoxing Huai. 2023. Aramus: Pushing the limits of data and model scale for arabic natural language processing. arXiv preprint arXiv:2306.06800.

Abbas Raza Ali, Muhammad Ajmal Siddiqui, Rema Algunaibet, and Hasan Raza Ali. 2023. A large and diverse arabic corpus for language modeling. arXiv preprint arXiv:2201.09227.

Ahmed Ali. 2018. Speech technology for arabic. In *Handbook of Speech Processing in Inflectional Languages*. Springer.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. arXiv preprint arXiv:2405.01590.

Al-Salman AbdulMalik Atwell Eric Alrabiah, Maha. 2014. The design and construction of the 50 million words ksucca king saud university corpus of classical arabic. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2)*, Lancaster University, UK.

Hana Alshaikh, Muath Nagi, et al. 2021. A survey of arabic language technologies in the big data era. *IEEE Access*.

Mohammed Altamimi and Abdulaziz M. Alayba. 2023. Anad: Arabic news article dataset. *Data in Brief*, 50:109460.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. 2024. Cidar: Culturally relevant instruction dataset for arabic. arXiv preprint arXiv:2403.17848.

Kareem Darwish. 2018. The challenges of social media analytics in the arabic language. In *Social Informatics*. Springer.

Ali El Filali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. https://huggingface.co/blog/leaderboard-arabic. Accessed: 2025-02-20.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2014. An arabic corpus of academic texts. In *Proceedings of LREC*.

Ibrahim Abu El-khair. 2016. 1.5 billion words arabic corpus. arXiv preprint arXiv:1611.04033.

Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Arabic text diacritization using deep neural networks. arXiv preprint arXiv:1905.01965.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4).

Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Philippe Langlais. 2022. Revisiting pre-trained language models and their evaluation for arabic natural language understanding. arXiv preprint arXiv:2205.10687.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick

Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang,

Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Mohamed Osman Hegazi, Yasser Al-Dossari, Abdullah Al-Yahy, Abdulaziz Al-Sumari, and Anwer Hilal. 2021. Preprocessing arabic text on social media. *Heliyon*, 7:e06191.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Ahmed Ibrahim, Muhammad Abdul-Mageed, et al. 2019. On dialectal variation in arabic social media. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*.

Abderrahmane Issam and Khalil Mrini. 2022. Goud.ma: A news article dataset for summarization in moroccan darija. In *Proceedings of the 3rd Workshop on African Natural Language Processing*.

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

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Annenberg School for Communication, University of Pennsylvania.

Imad Lakim, Ebtesam Almazrouei, Ibrahim Abualhaol, Merouane Debbah, and Julien Launay. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Mohammed Mghari, Omar Bouras, and Abdelaaziz El Hibaoui. 2022. Sanadset 650k: Data on hadith narrators. Mendeley Data.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, and Hasan Cavusoglu. 2020. Machine generation and detection of arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain.

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. arXiv preprint arXiv:2309.09400.

Rasha Obeidat, Marwa Al-Harbi, Mahmoud Al-Ayyoub, and Luay Alawneh. 2024. Arquad: An expert-annotated arabic machine reading comprehension dataset. *Cognitive Computation*, 16:1–20.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli,

Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.

Sebastian Ruder. 2018. Nlp's imagenet moment has arrived. The Gradient.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, and et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149.

Alex Wang, Amanpreet Singh, et al. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Asma Yamani, Raghad Alziyady, Reem AlYami, Salma Albelali, Leina Albelali, Jawharah Almulhim, Amjad Alsulami, Motaz Alfarraj, and Rabeah Al-Zaidy. 2024. The KIND dataset: A social collaboration approach for nuanced dialect data collection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, St. Julian's, Malta. Association for Computational Linguistics.

Waleed A. Yousef, Omar M. Ibrahime, Taha M. Madbouly, and Moustafa A. Mahmoud. 2019. Learning meters of arabic and english poems with recurrent neural networks: a step forward for language understanding and synthesis. *arXiv preprint arXiv:1905.05700*.

Wajdi Zaghouani. 2014. Critical survey of the freely available arabic corpora. In *Proceedings of LREC*.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus - preparation and integration into the clarin-infrastructure. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 175-182). Association for Computational Linguistics.

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11:147–151.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The nordic pile: A 1.2tb nordic dataset for language modeling. arXiv preprint arXiv:2303.17183.

10. Language Resource References

Al-Debsi, Ridhwan and Elnagar, Ashraf and Einea, Omar. 2019. *NADIA: News Articles Dataset in Arabic for Multi-Label Text Categorization*. Mendeley Data.

Al-Dulaimi, Ahmed Hashim. 2022. *Ultimate Arabic News Dataset*. Mendeley Data.

Alhagri, M. 2015. *Saudi Newspapers Arabic Corpus (SaudiNewsNet)*. GitHub.

ARBML Team. 2024. *Ashaar*. HuggingFace.

Boulahia, Ahmed Khalil. 2023. *Arabic Quotes Dataset*. GitHub. Licensed under Apache License 2.0.

Wikipedia Contributors. 2024. *Arabic Wikipedia Dump*. Wikimedia Foundation. Retrieved on February 2025.

Einea, Omar and Elnagar, Ashraf and Al-Debsi, Ridhwan. 2019. *SANAD: Single-Label Arabic News Articles Dataset for Automatic Text Categorization*. Mendeley Data.

El-Haj, Mahmoud and Koulali, Rim. 2013. *KALIMAT: A Multipurpose Arabic Corpus*. Semantic Scholar.

El Karef, Nehal. 2022. *Classical Arabic Poetry Dataset*. HuggingFace.

Elfilali, Ali. 2023. *Hindawi Books Dataset*. HuggingFace.

Farhat, May and Taghadouini, Said and Hallström, Oskar and Hajri-Gabouj, Sonja. 2024. *ArabicWeb24: Creating a High Quality Arabic Web-only Pre-training Dataset*. LightOn, INSAT.

Khalil, Ashwaq and Jarrah, Moath and Aldwairi, Monther. 2022. *Arabic Fake News Dataset (AFND)*. Mendeley Data.

Mohamed. 2018. *DataSet for Arabic Classification*. Mendeley Data.

Rezk, Abdelrahman. 2023. *Arabic Poem Comprehensive Dataset (APCD)*. HuggingFace.

Saad, Motaz. 2019. *Arabic-News*. GitHub.

Saad, Motaz and Ashour, Mohammed. 2019. *Arabic Billion Words Corpus*. GitHub. Large-scale Arabic corpus for NLP tasks.

Wenzek, Guillaume and Lachaux, Marie-Anne and Conneau, Alexis and others. 2020. *OPUS Wikipedia: A Large Multilingual Corpus of Wikipedia Articles*. OPUS / NLPL.

Zhang, Biao and Williams, Philip and Titov, Ivan and Sennrich, Rico. 2020. *OPUS-100: A Dataset for Multilingual Machine Translation*. OPUS / NLPL.

A. Dataset Card

Dataset Card for DIA2 Arabic Dataset

Dataset Name:
DIA2: A Comprehensive and Diverse Diacritized Modern Standard Arabic Corpus for Large-Scale NLP Research

Language: *Modern Standard Arabic (MSA) and Classical Arabic (CA)*

Size: 41.9 billion tokens

Key Features:

- Data and Code are publicly available
- Fully automatically diacritized for phonetic and syntactic tasks
- Gold diacritized subset available
- No translated or machine-generated content
- Compiled from credible sources
- Comprehensive cleaning (Preprocessing, URL filtering, deduplication)

Intended Use:

- Language modeling
- Text-to-speech
- MSA/Dialect classification
- Sentence completion
- Diacritics restoration

Limitations:

- No dialectal texts included
- Scarcity of clean gold diacritized data
- No sentence-level deduplication

Citation:

```
@inproceedings{dekmak2026dia2,
    title     = {DIA2: A Comprehensive and Diverse
    Diacritized Modern Standard Arabic Corpus for Large-
    Scale NLP Research},
    author    = {Dekmak et al},
    booktitle = {Proceedings of the OSACT7 Workshop at
    LREC-COLING 2026},
    year      = {2026}
}
```

B. Deduplication Parameters - Experiment

A mini-experiment was conducted on a sample of 682,004 documents, testing values of 20 and 32 for the number of permutations, and 0.4, 0.5, and 0.8 for the Jaccard similarity threshold. Prior to duplicate identification, 10 known duplicates with specific IDs were injected into the sample, leading to a total of 682,014 documents. This setup allowed us to evaluate the recall of each experiment.

For each deduplication test, we checked how many of these injected duplicates were identified (recall). We also inspected 10 identified duplicates, uniquely identified by each test, to determine if the results were actual duplicates or false positives (precision). The results of the precision and recall tests are presented in Table 7. Precision refers to the proportion of actual duplicates among the identified duplicates. Recall refers to how many of the 10 injected near-duplicates were identified. Given the results in 7, it is noticed that a lower

Permutations	Threshold	Recall	Precision	F1 score
20	0.4	1	0	0
20	0.5	1	0.4	0.57
20	0.8	0.1	1	0.18
32	0.4	1	0	0
32	**0.5**	1	1	**1**
32	0.8	0.1	1	0.18

Table 7: Precision and Recall: Precision is the number of actual duplicates identified out of 10 inspected duplicates, and Recall is the number of injected near-duplicates identified out of 10.

threshold (0.4, 0.5) allows more variation between documents before they are flagged as duplicates, which can increase recall (finding more duplicates) but may lead to false positives (incorrectly identifying documents as duplicates). Conversely, a higher threshold (0.8) requires documents to be more similar, which increases precision (correctly identifying true duplicates) but may reduce recall, as some near-duplicates may be missed. We also observed that using 32 permutations consistently provided more accurate and reliable results compared to 20 permutations.

Based on these results, the number of permutations was set to 32 and the Jaccard similarity threshold to 0.5. This balance provided the best trade-off between identifying duplicates and minimizing false positives, given the available computational resources.