ArabicWeb-Edu: Educational Quality Data for Arabic LLM Training

Majd Hawasly, Tasnim Mohiuddin, Hamdy Mubarak, Sabri Boughorbel
Qatar Computing Research Institute, HBKU
Doha, Qatar
{mhawasly,mmohiuddin,hmubarak}@hbku.edu.qa

Abstract

The quality of training data plays a critical role in the performance of large language models (LLMs). This is especially true for low-resource languages where high-quality content is relatively scarce. Inspired by the success of FineWeb-Edu (Lozhkov et al., 2024) for English, we construct a native Arabic educational-quality dataset using similar methodological principles. We begin by sampling 1 million Arabic web documents from Common Crawl and labeling them into six quality classes (0–5) with Qwen-2.5-72B-Instruct model using a classification prompt adapted from FineWeb-Edu. These labeled examples are used to train a robust classifier capable of distinguishing educational content from general web text. We train a classification head on top of a multilingual 300M encoder model, then use this classifier to filter a large Arabic web corpus, discarding documents with low educational value. To evaluate the impact of this curation, we pretrain from scratch two bilingual English-Arabic 7B LLMs on 800 billion tokens using the filtered and unfiltered data and compare their performance across a suite of benchmarks. Our results show a significant improvement when using the filtered educational dataset, validating the effectiveness of quality filtering as a component in a balanced data mixture for Arabic LLM development. This work addresses the scarcity of high-quality Arabic training data and offers a scalable methodology for curating educational quality content in low-resource languages.

1 Introduction

The remarkable progress of large language models (LLMs) in recent years has been fueled by the availability of massive and high-quality textual datasets (Soldaini et al., 2024). The quality of the underlying training data has emerged as a crucial factor in determining LLM performance, particularly in knowledge-intensive and instruction-following tasks (Rae et al., 2021; Groeneveld et al., 2024b; Wang et al., 2025). While significant advancements have been achieved for high-resource languages such as English, developing competitive LLMs for low-resource languages remains a substantial challenge. One of the key obstacles is the scarcity of curated, high-quality training corpora (Kreutzer et al., 2022), which limits the ability of LLMs to acquire rich linguistic, cultural, and domain-specific knowledge for these languages.

Arabic, spoken by over 400 million people across the globe, exemplifies this challenge. Despite its status as one of the most widely spoken languages, Arabic remains significantly underrepresented in existing LLMs, both in terms of training data coverage and downstream performance (Koto et al., 2024; Elfilali et al., 2024). A major bottleneck is the limited availability of large-scale, diverse, and high-quality Arabic textual resources. Existing Arabic web corpora often contain noisy, low-quality, or repetitive content, impeding the development of competitive Arabic LLMs.

Recent work on English LLM training has demonstrated that filtering web-scale corpora based on content quality can substantially improve model performance (Abdin et al., 2024; Penedo et al., 2024a). In particular, FineWeb-Edu (Lozhkov et al., 2024) introduced a methodology for curating English web data with a focus on educational value, yielding measurable improvements in downstream tasks. However, similar large-scale, quality-filtered datasets for Arabic are currently lacking, hindering

Proceedings of The Third Arabic Natural Language Processing Conference, pages 436–447

progress in building capable Arabic LLMs.

In this work, we introduce **ArabicWeb-Edu**, the first large-scale, educational-quality dataset designed specifically for training Arabic LLMs. Our approach systematically adapts and extends the quality-filtering principles established by FineWeb-Edu to the Arabic language context. We begin by sampling 1 million Arabic web documents from Common Crawl and employ a prompt-based classification strategy to assign content quality scores ranging from 0 to 5. These labeled examples are used to train a robust Arabic content quality classifier, enabling scalable filtering of large Arabic web corpora. We apply this classifier to filter out documents with low educational value (classes 0 and 1), thereby constructing a high-quality Arabic corpus focused on educational, informative, and linguistically rich content.

To assess the impact of our quality-filtering methodology, we pretrain bilingual English-Arabic LLMs on both the filtered and unfiltered datasets, totaling 800 billion tokens. We evaluate these models across a comprehensive suite of benchmarks. Our results demonstrate that models trained on the ArabicWeb-Edu dataset exhibit substantial improvements over their counterparts trained on unfiltered data, underscoring the critical role of content quality in enhancing LLM performance for Arabic. Our contributions are threefold:

- We construct ArabicWeb-Edu, a large-scale, educational-quality Arabic web corpus curated using a scalable, filtering methodology.
- We develop a robust and light-weight Arabic content quality classifier based on a multilingual encoder embedding model, facilitating reproducible and scalable filtering of Arabic web data.
- We provide empirical evidence, through rigorous LLM pretraining and evaluation, that quality-filtered Arabic data significantly enhances LLM performance across diverse benchmarks.

By addressing the long-standing scarcity of high-quality Arabic training data, this work contributes towards more equitable LLM development and highlights a scalable methodology for curating educational content in low-resource languages. We believe ArabicWeb-Edu when released will serve as a valuable resource for the community and a foundation for further advancements in Arabic LLM research.

2 Related Work

2.1 Arabic web data pipelines

ArabicWeb24 (Farhat et al., 2024) extracted Arabic documents from a custom 6.5 TB web crawl. Then, a `datatrove` (Penedo et al., 2024b)-based pipeline for filtration and deduplication was developed. The filtration concentrated on long or non-text documents, bad URLs of adult content, foreign languages, documents with unsuitable statistics, HTML elements and web page artifacts, and banned words. The pipeline resulted in a dataset of 28 billion tokens. Previous efforts include ArabicWeb16 (Suwaileh et al., 2016) which offered 10.8 TB data from 150M Arabic web pages.

The multilingual OSCAR project (Ortiz Suárez et al., 2019) also offers a filtered collection of Arabic web data using high-performance data pipelines with a special focus on data quality. The latest version of the corpus (23.01) offers 10 billion words from 25M documents. Another multilingual effort is arTenTen (Arts et al., 2014) from the TenTen corpus family offering 6.5 billion words in its most recent release.

2.2 FineWeb-Edu dataset

FineWeb (Penedo et al., 2024a) is a 15-trillion-token English dataset derived from 96 Common Crawl snapshots, processed through a sophisticated pipeline involving filtering and deduplication. **FineWeb-Edu** (Lozhkov et al., 2024) is a 1.3-trillion token subset of FineWeb, extracted using an educational quality classifier. To train this classifier, `LLaMA3-70B-Instruct` was used to label 500k FineWeb samples with an educational score ranging from 0 to 5, where 0 denotes no educational value and 5 indicates high-quality educational content. A BERT-style regression model was then fine-tuned on this labeled data. Finally, the full FineWeb dataset was scored using the trained classifier, and only documents with a score of 3 or higher were retained to form FineWeb-Edu.

Inspired by its success, Alrashed et al. (2024) translated the deduplicated version of Fineweb-Edu from English to Arabic in the training split of `SmolLM` model with `nllb-200-distilled-600M`. Also, (Yu et al., 2025) is a Chinese adaptation of the FineWeb-Edu approach to Chinese content with total 1.5T tokens in the v2.1 release.

Edu class	0	1	2	3	4	5
Ratio %	1.3	24.5	50.8	20.1	3.1	0.02

Table 1: Seed dataset's education class distribution

Edu class	Precision	Recall	F1	Support%
0	0.71	0.14	0.24	1.3
1	0.65	0.44	0.53	24.8
2	0.63	0.83	0.72	50.6
3	0.60	0.47	0.53	20.1
4	0.64	0.18	0.29	3.1
5	0.00	0.00	0.00	0.024
Avg. macro	0.54	0.35	0.38	
Avg. weighted	0.63	0.63	0.61	

Table 2: Classifier validation on the test set of size 100k Arabic web documents.

3 ArabicWeb-Edu Construction

To construct an Arabic web corpus enriched with high-quality educational content, we adopt the scalable methodology inspired by FineWeb-Edu (Lozhkov et al., 2024), with careful adaptations to address the linguistic and resource-specific challenges of Arabic. Our approach consists of three key stages: (i) labeling a high-quality seed dataset of Arabic web documents with educational quality scores, (ii) training a robust classifier to scale this annotation to large web corpora, and (iii) large-scale corpus filtering.

3.1 Seed dataset labeling

The first stage involves building a high-quality, labeled dataset to serve as the foundation for classifier training. We randomly sampled 1 million Arabic web documents from recent Common Crawl snapshots, ensuring a diverse representation of Arabic web content across topics and domains. To annotate these documents with educational quality labels, we leverage `Qwen-2.5-72B-Instruct` (Qwen Team, 2024) due to its strong performance in Arabic[1] and extensive context length. A tailored zero-shot prompt was used to define and distinguish the six levels of educational quality, ranging from 0 (lowest quality, no educational value of any kind) to 5 (highest quality, content suited for teaching); the prompt could be found in Appendix B. Using this prompt, each sampled document is scored independently by the model, resulting in a quality-labeled seed dataset of 1 million Arabic web documents[2]. Table 1 presents the class distribution of the labeled seed dataset, illustrating the relative prevalence of different quality levels in the Arabic web domain.

3.2 Educational quality classifier training

Using the labeled seed dataset, we train a dedicated document-level classifier to automatically predict the educational quality of Arabic web documents at scale. We adopt the `mGTE` architecture (Zhang et al., 2024) for this task—a 305M parameter multilingual encoder with an 8k token context window. This architecture offers an effective balance between model capacity and computational efficiency, enabling scalable document-level classification without sacrificing performance on long-context inputs, which are common in web data. We train a multi-class classifier to predict the educational quality score (0–5). Specifically, we fine-tune the pretrained `mGTE-305M` model on the 900k training split of the labeled data for 20 epochs with a learning rate of $3e^{-4}$. To preserve the model's general linguistic capabilities while specializing it for the classification task, we freeze the embedding and encoder layers, updating only the task-specific classification head. We selected the checkpoint with the highest F1 score on the held-out validation set. The accuracy of the trained classifier is 0.63, likely due to the natural distribution of Arabic web documents that lacks high quality content.

Table 2 shows the precision, recall and F1 scores for the classifier on the test split of 100k documents. The confusion matrix can be seen in Appendix A.

3.3 Large-scale corpus filtering

The trained classifier is applied to a large Arabic web corpus, enabling systematic filtering based on educational quality. Through empirical analysis, we define documents with quality scores 0 or 1 as having low educational value[3] and exclude them from the final corpus. The remaining documents, which span quality classes 2 to 5, constitute the educationally filtered Arabic web corpus suitable for LLM pretraining. This scalable filtering process enables the construction of an Arabic web corpus with significantly enriched educational content, addressing the long-standing scarcity of high-quality Arabic resources for LLM development.

[1] hf.co/spaces/OALL/Open-Arabic-LLM-Leaderboard

[2] The 1M document seed dataset is released at hf.co/datasets/sboughorbel/arabic-web-edu-seed

[3] Sample documents from the different educational classes can be seen in Appendix C

Model	MMMLU/Ar (0-shot)	ArabicMMLU (3-shot)	ACVA (5-shot)	PIQA/MSA (0-shot)	OALL-v1 (0-shot)	OALL-v2 (0-shot)
Baseline@826B	23.47	31.67	49.10	**62.62**	34.50	31.50
Edu@841B	**24.26**	**32.45**	**55.28**	61.64	**36.93**	**34.69**
Change	*+3.37%*	*+2.46%*	*+12.59%*	*-1.56%*	*+7.04%*	*+10.13%*

Table 3: Modern Standard Arabic (MSA) benchmarking results.

Model	Belebele/Ar (3-shot)	PIQA/Egy (0-shot)	PIQA/Lev (0-shot)	ArabicMMLU/Egy (0-shot)	ArabicMMLU/Lev (0-shot)
Baseline@826B	**26.80**	**59.41**	**56.64**	27.61	28.17
Edu@841B	26.41	58.87	54.57	**29.12**	**34.59**
Change	*-1.45%*	*-0.91%*	*-3.65%*	*+5.47%*	*+22.79%*

Table 4: Dialectal benchmarking results.

4 Empirical Evaluation

To empirically evaluate the impact of our dataset on LLM pretraining, we conducted an ablation study: we trained from scratch a baseline model with the `OLMo-7B` architecture (Groeneveld et al., 2024a) on a balanced mixture of Arabic, English and code data, derived from the data mix of Fanar suite of models (Fanar Team et al., 2025). We compare this model with an identical setup in which the web portion of the Arabic data is replaced with our ArabicWeb-Edu dataset. We benchmark the closest two checkpoints to the 800B token point on a suite of standard evaluation tasks to assess performance differences. The tasks are:

- MMMLU: the Arabic subset of OpenAI's professionally translated MMLU dataset (Hendrycks et al., 2021).
- ArabicMMLU (Koto et al., 2024): a multi-choice dataset of Arabic knowledge.
- ACVA (Huang et al., 2024): the Arabic Cultural & Value Alignment dataset.
- OALL (Elfilali et al., 2024): a suite of varied Arabic language understanding tasks. We show both versions of this benchmark.
- AraDiCE (Mousi et al., 2025): a suite of professionally translated subsets of PIQA and ArabicMMLU datasets to dialectal Egyptian and Levantine.
- Belebele (Bandarkar et al., 2024): the average of the six Arabic dialects from Belebele (namely, `acm_Arab`, `apc_Arab`, `arb_Arab`, `ars_Arab`, `ary_Arab` and `arz_Arab`).

Table 3 presents the benchmarking results for MSA tasks. The results show a notable improvement on almost all tasks. The holistic OALL benchmark especially shows a significant jump.

In contrast, for dialectal benchmarking (Table 4) regression could be observed in some benchmarks. This could be a direct result of losing dialectal web content due to rigorous educational filtering. Thus, we see this approach as a component in a balanced data mix strategy that augments the filtered web content with better quality data extracted from books and trusted sources, in addition to other data that does not qualify as educational but are important for training, including dialogue and dialectal content.

5 Conclusion

This work demonstrates that quality-based data curation significantly enhances the performance of low-resource language models, addressing a critical challenge in Arabic LLM development. Our work makes two key contributions to the field: first, it provides a scalable solution to the scarcity of high-quality Arabic training data, and second, it establishes a replicable methodology that can be extended to other low-resource languages.

The success of this approach suggests that investment in careful Arabic data curation can yield significant returns in model performance, offering a practical path forward for developing more capable language models across diverse linguistic contexts. Future work would investigate the extension of quality filtering to better handle Arabic dialectal content.

Limitations

The rigorous educational filtering process appears to disproportionately remove dialectal Arabic content, as evidenced by performance regression on dialectal benchmarks (Table 4). This limitation restricts the model's ability to understand and generate content in Arabic dialects, potentially limiting its applicability for diverse Arabic-speaking populations.

- Limited Domain Coverage: The focus on educational content may inadvertently bias the dataset toward formal, academic domains while underrepresenting other valuable linguistic patterns and cultural expressions present in everyday Arabic web content. This could impact the model's performance on creative, conversational, or culturally-specific tasks.
- Evaluation Scope: In this short paper, our empirical evaluation is limited to a 7B parameter model architecture and specific benchmark tasks. The generalizability of these findings to larger models, different architectures, or alternative evaluation metrics remains to be validated. Additionally, the evaluation focuses primarily on knowledge-intensive tasks, which may favor educational content filtering but not reflect performance on other important capabilities.
- Scalability and Computational Requirements The two-stage filtering process (LLM annotation followed by classifier training) requires significant computational resources and may not be easily replicable for researchers with limited access to large language models. The reliance on Qwen-2.5-72B for initial labeling creates a dependency on proprietary models.
- Cultural and Linguistic Bias: The educational quality criteria adapted from FineWeb-Edu were originally designed for English content and may not fully capture the educational value standards appropriate for Arabic content across different cultural contexts within the Arabic-speaking world.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Sultan Alrashed, Dmitrii Khizbullin, and David R Pugh. 2024. Fineweb-Edu-Ar: Machine-translated corpus to support Arabic small language models. *arXiv preprint arXiv:2411.06402*.

Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, and Vit Suchomel. 2014. artenten: Arabic corpus and word sketches. *Journal of King Saud University - Computer and Information Sciences*, 26(4):357–371. Special Issue on Arabic NLP.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ali Elfilali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open Arabic llm leaderboard. `https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard-v1`.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

May Farhat, Said Taghadouini, Oskar Hallström, and Sonja Hajri-Gabouj. 2024. ArabicWeb24: Creating a high quality arabic web-only pre-training dataset.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 22 others. 2024a. OLMo: Accelerating the science of language models. *Preprint*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024b. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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, and 33 others. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu: the finest collection of educational content.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *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 allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. Datatrove: large scale data processing.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Reem Suwaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. ArabicWeb16: A new crawl for today's Arabic web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 673–676, New York, NY, USA. Association for Computing Machinery.

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, and 1 others. 2025. Ultrafineweb: Efficient data filtering and verification for high-quality llm training data. *arXiv preprint arXiv:2505.05427*.

Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. 2025. Opencsg chinese corpus: A series of high-quality chinese datasets for llm training. *Preprint*, arXiv:2501.08197.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

A Confusion Matrix

Figure 1 shows the normalized confusion matrix for the ArabicWeb-Edu classifier.

B Seed Classification Prompt

The box shows the prompt used with `Qwen2.5-72B-Instruct` to create the seed

Prompt

Below is an extract in Arabic from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
- Grant a fourth point if the extract highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.

The extract:
<EXAMPLE>.

After examining the extract:

- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

C Samples of Educational Classes

Data samples from the educational classes are presented in Tables 5 and 6. English translations of the same samples can be found in Tables 7 and 8.

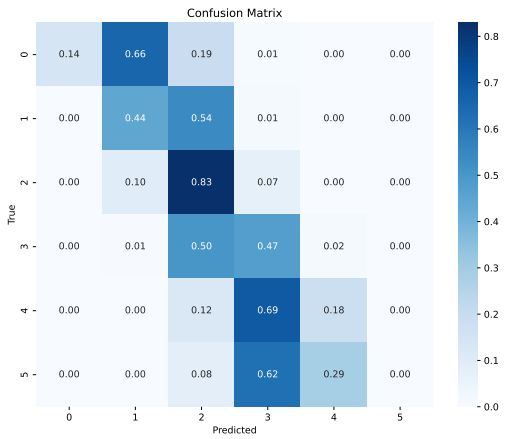

Figure 1: Classifier normalized confusion matrix

Class 0 (mostly harmful/explicit content)	تفاصيل الاعتداء على مثلي جنسيا بتريبك و الجناة بينهم فتاة قاصر! ... ووفقا لمصادر إعلامية محلية ، فالجناة ٣ منهم يتابعون دراستهم بمعهد التكوين المهني بتريبك، وفتاة قاصر لا يتجاوز عمرها ١٦ سنة تتابع دراستها بإحدى الثانويات بنفس المنطقة، أجبروا الضحية العاطل عن العمل على مرافقتهم إلى جبل، و هناك جردوه من ملابسه، واعتدوا عليه بشكل شنيع، ما أدى إلى إصابته إصابات بليغة، دفعته إلى اللجوء إلى الشرطة بولاية مسريف، حيث وضع شكاية بالمتهمين.
	أفضل الكازينوهات على الإنترنت للاعبين في العالم العربي يبحث محبي العاب المراهنات عبر الإنترنت عن أفضل العاب الكازينو اون لاين. يمكنك قضاء بعض الوقت في الاستمتاع بهذه الألعاب بعد يوم طويل وشاق من العمل، أو إذا كنت تشعر ببعض الملل. فألعاب الكازينو اون لاين تتميز بالإثارة والتشويق لدرجة أنه قد تنسى اوقات الانتظار الملة من خلال هذه الألعاب. لم تعد هناك مشكلة في الوصول إلى هذه الألعاب كما في الماضي. أفضل مواقع العاب كازينو اون لاين لشهر ٠٥ ٢٠٢١ ...
Class 1	صنع لجميع المصانع و المستوردين في مصر - صنع لجميع المصانع و المستوردين في مصر كل الضرائب مضافة لسعر المنتج الشحن عند متابعة عملية الشراء. يحظر لي عندما يكون هذا المنتج هو متاح: الرجاء إبلاغي عندما يتوفر عطر Hypnotic Poison Eau Sensuelle Christian Dior للنساء ([البديل]) - ده الاصدار الرسمي الاصلي المبرشم و القادم من الماركات و الاعلي سعرا ده الاصلي الاصلي ايضا القادم في بوكس ابيض او بيج - مكتوب علي العلبة و الزجاجة اه تستر و ليس للبيع - بلا عيب قادم من الماركات - سعره اقل قليلا من الماستر العطر نفس جودة النوعين السابقيين من كل شيء (رائحة - ثبات - فوحان) - بدون علب - بالزجاجة عيوب شحن خارجية و عيوب تخزين حيث انه تقليد للعطر الاصلي بالعلبه التستر بنسبة ٩٠ - ٩٥% من حيث الشكل و الرائحة و الثبات و الفوحان ...
	... فئة المنتج من دواسات قابلة للطي، ونحن المصنعين المتخصصة من الصين، دواسات قابلة للطي، قطع غيار دراجات المحيط الهادئ الموردين مصنع بالجملة منتجات ذات جودة عالية من مكونات دراجات Aest R & D والتصنيع، لدينا الكمال خدمة والدعم الفني ما بعد البيع. نتطلع الى تعاونكم! ...
Class 2	صدر عن مركز الأدب العربي للنشر والتوزيع، كتاب نصوص بعنوان (موسم القطاف لامرأة من خريف) للكاتبة أماني ظافر. (موسم القطاف لامرأة من خريف) هو نصوص أدبية لامرأة عربية تروي واقعها في أسطر قليلة حمّلتها أمانة الوصول دون مبالغة، صفحات يكسوها الصمت وتملأها لغة النساء وحدها، فهي نصوص لا تربطها علاقة عدا أنها تشاركت ذنب الأقدام والمرأة ذاتها. ويتساءل الكتاب: ما الذي يجعل امرأة عربية دون حصانة تكتب؟، تنصهر في كلمات وتهدهد الحقائق دون أجر!، امرأة بمتلازمة أحلام لا حدود لها وواقع متقزم جداً لا يتسع لكليهما معاً!، اندثرت جُل احلامها كمذكرات سجين لن يخرج من زنزانته بشيء عدا وباء ذاكرة وبداية تائهة وسجل سوابق زخماً!، امرأة عربية بخطا اثقل من ذوات جنسها في العالم ، يتقدمها الذكور بخطوات شاسعة دون عدالة شارة البدء، لم يكن الرهان على اللياقة، كان خط البداية غير منصف، فبدء انطلاقهما كان متأخراً جداً في حين أن الذكور كانوا قد قطعوا ميلاً من الحياة! ... .
	إحباط محاولة أقتحام على السفارة الروسية في كابول حسبما ذكر قناة العالم ينقل لكم موقع صحيفة الوسط محتوي خبر إحباط محاولة أقتحام على السفارة الروسية في كابول. وتم اِلْتِقَاط بالقرب من مبنى البعثة الدبلوماسية الروسية على سيارة مازدا، ووفقا للبيانات الأولية، فإن السيارة تحتوي على ١٠٠٠ كيلوغرام من متفجرات تي إن تي. وفي الوقت نفسه، هناك معلومات تفيد بأن هناك أقتحام قد حصل باستخدام السيارة المفخخة بالقرب من السفارة. ولا يوجد تأكيد رسمي على المعلومات المتعلقة بمحاولة الهجوم بعد.

Table 5: Samples of educational quality classes. The examples of class 0 were particularly cherry-picked not to offend or harm the readers as the class covers mostly very harmful and explicit content.

Class 3	هل تدري أن أصعب معركة هي معركة الانتصار على الذات أقوال و حكم بالعربي مَن يهزم رغباته أشجع ممن يهزم أعداءه، لأن أصعب انتصار هو الانتصار على الذات. — أرسطو إن الإنسان لن يملُكْ السعادَة إلا إذا طوّر مَلَكاتِه وقُدراتِه . — أرسطو إننا لا نصطاد الثعلب بالفخ نفسه مرتين. — أرسطو مَن يهزم رغباته أشجع ممن يهزم أعداءه، لأن أصعب انتصار هو الانتصار على الذات. — أرسطو كان هيجل على حق عندما قال أننا نتعلم من التاريخ أنه يستحيل على البشر التعلم من التاريخ. من عرف نفسه لا يضره ما يقوله الناس فيه. وإذا عرفت هفوة مسلم .. فانصحه بالسر .. وإذا وعظته، فلا تعظه وأنت مسرور باطلاعك على نقصه لينظر إليك بعين التعظيم، وتنظر إليه بعين الاستحقار، وتترفع عليه بدالة الوعظ، وليكن قصدك تخليصه من الإثم وأنت حزين، كما تحزن على نفسك إذا دخل عليك نقصان في دينك. وينبغي أن يكون تركه لذلك من غير نصحك أحب إليك من تركه بالنصيحة، فإذا فعلت ذلك، كنت قد جمعت بين أجر الوعظ وأجر الغم بمصيبته وأجر الإعانة له على دينه.
	الرئيسية / ارشادات / كيف تجعل الايفون ينطق اسم المتصل عند المكالمات؟ كيف تجعل الايفون ينطق اسم المتصل عند المكالمات؟ يحتوي هاتف الايفون على ميزة هامة للغاية وهي نطق اسم المتصل عند ورود مكالمة جديدة، وعلى الرغم أن الميزة موجودة منذ سنوات ألا أن الكثير من المستخدمين يجهل وجودها أو كيفية تفعيلها. هناك عدة أسباب تجعل ميزة نطق اسم المتصل في الايفون مفيدة مثل أن يكون الهاتف بعيداً عنك أو موضوعاً في جيبك، أو أثناء قيادة السيارة حتى لا يتشتت ذهنك بالنظر إلى شاشة الهاتف لمعرفة هوية المتصل، أو عند ارتداء سماعات البلوتوث، كل هذه المواقف وأكثر سوف تكون فيها تلك الخاصية على الايفون ذات فائدة عظيمة. ...
Class 4	مع مرور الأيام تزداد سرعة إنتشار وقوة فيروس كورونا المستجد، هذا المرض الذي أصبح يهدد الكثيرين في مختلف دول العالم، ينتقل بطرق مختلفة سواء من خلال العطس أو اللمس، ما يجعل من الصعب جدّاً الحدّ منه. وهنا نشير الى أن هناك العديد من الخطوات والإجراءات الوقائية التي تساهم في منع الإصابة بهذا النوع من العدوى، لا سيما خلال التواجد في مكان العمل خلال التواجد في أماكن العمل، من الضروري الإلتزام بالعديد من الخطوات الأساسية التي لها دور أساسي في الوقاية من خطر إنتقال مرض الكورونا، ومن أهمها: الحرص على غسل اليدين جيداً بالماء والصابون وذلك لمرّات متكررة في اليوم الواحد، ما يعتبر وسيلة أساسية للتخلّص من تراكم الجراثيم والبكتيريا الضارّة. لا بدّ من تفادي لمس العيون والأنف والفم بأيدي غير مغسولة بشكل جيّد، حيث أن ذلك يساعد في إنتشار الفيروس بشكل أكبر. من الضروري تجنّب إجراء أيّ إتصال مباشر مع المحيطين بك في العمل، والإمتناع تماماً عن العناق والتقبيل. لحماية جهازك التنفسي من العدوى، لا تتردد بإرتداء الكمامة طوال فترة تواجدكَ في مكان العمل. لا يجب الإغفال عن تنظيف وتطهير أسطح المكاتب وأجهزة الكمبيوتر والهواتف المستعملة في مكان الوظيفة إضافةً الى مقابض الأبواب، حيث أنها من الممكن أن تكون حاملة للفيروس بشكل كبير.
	طقس العرب — أكد علماء في جامعة سوانسي وهيئة المساحة البريطانية للقارة القطبية الجنوبية أن واحدا من أضخم جبال الجليد على الإطلاق انفصل عن القارة القطبية الجنوبية ليشكل مخاطر على السفن أثناء تفتته. وأوضحوا أن الجبل الذي يزن نحو تريليون طن وحجمه ٥٨٠٠ كيلومتر مكعب انفصل عن الحرف الجليدي (لارسن سي) في القارة القطبية الجنوبية في الفترة الممتدة بين ١٠ إلى ١٢ يوليو تموز، وفقا لـ رويترز. وبين الاستاذ بجامعة سوانسي والمحقق الرئيسي في مشروع ميداس الذي يراقب الحرف الجليدي منذ سنوات أدريان لوكمان، أن جبل الجليد واحد من أكبر جبال الجليد التي جرى رصدها ومن الصعب التنبؤ بتطوره المستقبلي ...

Table 6: Samples of educational quality classes.

Class 0	Details of the assault on a homosexual in Trebek... and the perpetrators included an underage girl! According to local media sources, the perpetrators include three students at the Trebek Vocational Training Institute, and a 16-year-old girl studying at a high school in the same area. They forced the unemployed victim to accompany them to a mountain, where they stripped him of his clothes and brutally assaulted him, causing him severe injuries. He was then forced to go to the police in Mesrif, where he filed a complaint against the accused.
	The Best Online Casinos for Players in the Arab World Online gambling enthusiasts are looking for the best online casino games. You can spend some time enjoying these games after a long, tiring day at work, or if you're feeling a little bored. Online casino games are so exciting and thrilling that you can forget about those boring wait times. Accessing these games is no longer a problem like in the past. The Best Online Casino Sites for May 2021
Class 1	Made for all factories and importers in Egypt - Made for all factories and importers in Egypt All taxes are added to the product price and shipping upon completion of the purchase. Notify me when this product is available: Please notify me when the LR Hypnotic Poison Eau Sensuelle Christian Dior for women ([alternative]) perfume is available. This is the official, original, sealed version, coming from the brands and the highest price. This is the original, also original, coming in a white or beige box. It is written on the box and bottle: "Tester, not for sale." It is flawless and comes from the brands. Its price is slightly lower than the Master. The perfume is of the same quality as the previous two types in every way (scent, durability, and sillage). It is without a box. The bottle has shipping and storage defects, as it is a 90-95% imitation of the original perfume in the concealed box, in terms of appearance, scent, durability, and sillage. . . .
	Product category: Folding Pedals. We are manufacturers. Specializing in folding pedals and bicycle spare parts from China. Pacific Ocean Factory Suppliers, Wholesale High-Quality Products. From bicycle components to R&D and manufacturing, we offer perfect after-sales service and technical support. We look forward to your cooperation!
Class 2	The Arab Literature Center for Publishing and Distribution has published a book of texts titled "The Harvest Season of a Woman from Autumn" by Amani Dhafer. "The Harvest Season of a Woman from Autumn" is a collection of literary texts by an Arab woman who narrates her reality in a few lines, conveyed without exaggeration by the trust of access. Pages covered in silence and filled with the language of women alone, these texts are unrelated except for the fact that they share the guilt of feet and women themselves. The book asks: What makes an Arab woman without immunity write? Melt into words and soothe truths without compensation! A woman with a syndrome of limitless dreams and a very dwarfed reality that cannot accommodate both of them together! Most of her dreams have vanished like the memoirs of a prisoner who will not emerge from his cell with anything but an epidemic of memory, a lost beginning, and a record of momentum! An Arab woman with a footstep heavier than those of her gender in the world, preceded by men with great strides without the fairness of the starting signal. The bet was not on fitness; the starting line was unfair, as their departure was too late in While the males had already walked a mile of life!. . . .
	Foiled attack on Russian embassy in Kabul According to Al-Alam TV, Al-Wasat newspaper website reports that an attempted attack on the Russian embassy in Kabul has been thwarted. A Mazda vehicle was spotted near the Russian diplomatic mission building, and according to preliminary information, the vehicle contained 1,000 kilograms of TNT. At the same time, there is information indicating that a car bomb was used to storm the embassy. There is no official confirmation of the information regarding the attempted attack yet.

Table 7: English translation of the the samples of educational quality classes in Table 5. The examples of class 0 were particularly cherry-picked not to offend or harm the readers as the class covers mostly very harmful and explicit content.

Class 3	Did you know that the most difficult battle is the battle of self-defeat? Sayings and Proverbs in Arabic He who defeats his desires is braver than he who defeats his enemies, because the most difficult victory is the victory over the self. — Aristotle A person will not attain happiness unless he develops his faculties and abilities. — Aristotle We never catch the fox in the same trap twice. — Aristotle He who defeats his desires is braver than he who defeats his enemies, because the most difficult victory is the victory over the self. — Aristotle Hegel was right when he said that we learn from history that it is impossible for humans to learn from history. He who knows himself is not harmed by what people say about him. If you become aware of a Muslim's fault, then advise him in secret. If you preach to him, do not preach to him while you are happy to know of his shortcomings, lest he look at you with respect and you look at him with contempt, and act superior to him with preaching. Let your intention be to free him from the sin while you are sad, just as you would be sad for yourself if you saw a deficiency in your faith. Leaving him alone without advising him should be more beloved to you than leaving him alone with advice. If you do that, you will have combined the reward of preaching, the reward of being saddened by his affliction, and the reward of helping him in his faith.
	Home/Guidelines/How to make your iPhone speak the caller's name during calls? How to make your iPhone speak the caller's name during calls? The iPhone has a very important feature: speaking the caller's name when a new call comes in. Although this feature has been around for years, many users are unaware of its existence or how to activate it. There are several reasons why the iPhone's speaking caller name feature is useful, such as when your phone is far away or in your pocket, when driving so you don't get distracted by looking at the phone screen to see who's calling, or when wearing Bluetooth headphones. In all these situations and more, this feature on the iPhone will be of great benefit.
Class 4	As the days pass, the spread and severity of the novel coronavirus increases. This disease, which has become a threat to many people in various countries around the world, is transmitted in various ways, whether through sneezing or touch, making it extremely difficult to control. Here, we note that there are many preventative steps and measures that help prevent this type of infection, especially during work. While working in the workplace, it is essential to adhere to several basic steps that play a fundamental role in preventing the risk of transmission of the coronavirus. The most important of these are: – Ensure that hands are thoroughly washed with soap and water, several times a day, as this is an essential means of eliminating the accumulation of harmful germs and bacteria. – Avoid touching your eyes, nose, and mouth with unwashed hands, as this further contributes to the spread of the virus. – It is essential to avoid any direct contact with those around you at work, and to completely refrain from hugging and kissing. To protect your respiratory system from infection, don't hesitate to wear a mask throughout your time at work. Don't forget to clean and disinfect desk surfaces, computers, and phones used at work, as well as door handles, as they can be a significant carrier of the virus.
	Arab Weather — Scientists at Swansea University and the British Antarctic Survey have confirmed that one of the largest icebergs ever recorded has broken away from Antarctica, posing a risk to ships as it disintegrates. They explained that the iceberg, weighing about a trillion tons and measuring 5,800 cubic kilometers, broke away from the Larsen C ice shelf in Antarctica between July 10 and 12, according to Reuters. Adrian Luckman, a professor at Swansea University and principal investigator of Project MIDAS, which has been monitoring the ice shelf for years, said that the iceberg is one of the largest ever observed, and its future development is difficult to predict.. . .

Table 8: English translations of the samples of educational quality classes in Table 6.