AraModernBERT: Transtokenized Initialization and Long-Context Encoder Modeling for Arabic

Omar Elshehy[1,7] **Omer Nacar**[2,7] **Abdelbasset Djamai**[3,7] **Muhammed Ragab**[4,7]
Khloud Al Jallad[5,7] **Mona Abdelazim**[6,7]

[1]Universität des Saarlandes, [2]Tuwaiq Academy, [3]Datategy,
[4]Leibniz-Institute for Educational Media | Georg-Eckert-Institute,
[5]Arab International University, [6]Ain Shams University, [7]NAMAA Community
onajar@tuwaiq.edu.sa

Abstract

Encoder-only transformer models remain widely used for discriminative NLP tasks, yet recent architectural advances have largely focused on English. In this work, we present **AraModernBERT**, an adaptation of the ModernBERT encoder architecture to Arabic, and study the impact of transtokenized embedding initialization and native long-context modeling up to 8,192 tokens. We show that transtokenization is essential for Arabic language modeling, yielding dramatic improvements in masked language modeling performance compared to non-transtokenized initialization. We further demonstrate that AraModernBERT supports stable and effective long-context modeling, achieving improved intrinsic language modeling performance at extended sequence lengths. Downstream evaluations on Arabic natural language understanding tasks, including inference, offensive language detection, question-question similarity, and named entity recognition, confirm strong transfer to discriminative and sequence labeling settings. Our results highlight practical considerations for adapting modern encoder architectures to Arabic and other languages written in Arabic-derived scripts.

1 Introduction

Transformer-based encoder-only language models such as BERT have become essential components of modern natural language processing pipelines, especially for retrieval, classification, and representation learning tasks (de Vries and Nissim, 2021; Karpukhin et al., 2020; Khattab and Zaharia, 2020). Despite the recent dominance of large autoregressive language models, encoder-based architectures remain widely deployed due to their favorable trade-offs in efficiency, latency, and scalability. Recent work has significantly modernized encoder architectures through improved attention mechanisms, positional encodings, and hardware-aware design, leading to substantial gains in performance and efficiency (Warner et al., 2025). However, these advances have been developed and evaluated primarily for English, and their transfer to Arabic and other languages using the Arabic script remains comparatively underexplored.

Arabic presents distinct challenges for encoder-based modeling. Its rich and templatic morphology, high lexical sparsity, and orthographic variation amplify the importance of tokenizer design and embedding initialization strategies (Rust et al., 2021; Petrov et al., 2023). Multilingual and English-centric tokenizers often fragment Arabic words excessively, resulting in longer effective sequence lengths and poorly trained subword embeddings. These issues are further compounded in Arabic-language domains such as news, legal texts, religious writings, and encyclopedic content, where documents frequently exceed the 512-token context limit of classical BERT-style models (Antoun et al., 2020; Abdul-Mageed et al., 2021; Inoue et al., 2021). As a result, both tokenization quality and long-context modeling are particularly important for Arabic, yet their interaction with modern encoder architectures has not been systematically studied.

In this paper, we introduce **AraModernBERT**, an Arabic adaptation of the ModernBERT encoder architecture (Warner et al., 2025). Rather than proposing a new model family, we focus on carefully transferring a modernized encoder design to Arabic and empirically analyzing two key factors: *transtokenized embedding initialization* and *native long-context modeling up to 8,192 tokens*. Transtokenization aligns a newly trained tokenizer with pretrained representations by initializing target-language embeddings from semantically aligned

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 313–321
March 28, 2026.

source-language embeddings, thereby mitigating the mismatch between tokenizer vocabularies and embedding spaces (Remy et al., 2024). Long-context modeling, enabled by architectural design choices such as alternating local and global attention and rotary positional embeddings (Su et al., 2021), allows the encoder to process substantially longer sequences than traditional Arabic BERT variants.

We conduct a comprehensive evaluation spanning intrinsic language modeling, downstream Arabic natural language understanding (NLU) tasks, and retrieval. Our experiments show that transtokenization is essential for stable and effective Arabic encoder training, achieving considerable improvements in masked language modeling performance compared to non-transtokenized initialization. We further demonstrate that AraModernBERT supports stable long-context modeling, achieving improved masked language modeling performance at extended sequence lengths without numerical instability or excessive memory usage. Downstream evaluations on Arabic natural language understanding tasks, including natural language inference (NLI), offensive language detection, and question–question similarity, confirm strong transfer to discriminative settings (Antoun et al., 2020; Abdul-Mageed et al., 2021).

This work provides practical insights into adapting modern encoder architectures to Arabic. By focusing on tokenizer initialization and long-context modeling, we highlight design considerations that are broadly applicable to Arabic and other Arabic-script languages. We release AraModernBERT and our evaluation code to support further research in this space.

2 Related Work

Encoder-only transformer models have been widely adopted for Arabic NLP, with AraBERT and its variants establishing strong baselines for Modern Standard Arabic and selected dialects (Antoun et al., 2020). Subsequent work, including CAMeLBERT and MARBERT, demonstrated the importance of domain selection and dialectal coverage for Arabic pretraining (Inoue et al., 2021; Abdul-Mageed et al., 2021). Despite their effectiveness, these models largely inherit the original BERT design, including a fixed 512-token context limit and absolute positional embeddings, which restrict their applicability to long Arabic documents commonly found in news, legal, and religious domains.

While some recent Arabic encoder efforts focus on efficiency or specialization, architectural modernization has largely lagged behind advances developed for English-language encoders. In contrast, a growing body of work revisits encoder design more broadly. Models such as MosaicBERT, AcademicBERT, and CrammingBERT explore training efficiency and resource-constrained settings, but do not substantially alter core architectural assumptions such as context length or attention structure. More recent long-context encoders, including NomicBERT and GTE-en-MLM, extend sequence length primarily for retrieval-oriented applications, but are trained and evaluated almost exclusively on English, limiting their relevance to morphologically rich and under-resourced languages.

ModernBERT represents a significant step forward in encoder architecture by incorporating alternating local and global attention, rotary positional embeddings, and hardware-aware design, enabling native processing of sequences up to 8,192 tokens while maintaining high efficiency (Warner et al., 2025). Our work builds directly on this architecture and investigates its transfer to Arabic, a setting not explored in the original ModernBERT study.

Tokenization has been shown to play a central role in multilingual and low-resource language modeling. Prior work demonstrates that multilingual subword tokenizers disproportionately benefit high-resource languages with shared alphabets, often leading to excessive fragmentation and poorly trained embeddings for languages such as Arabic (Rust et al., 2021; Petrov et al., 2023). Vocabulary transfer has therefore emerged as a promising strategy for language adaptation, with early approaches relying on embedding alignment or token reuse based on orthographic similarity (Artetxe et al., 2020; de Vries and Nissim, 2021). However, these methods are limited by tokenizer overlap and language proximity.

Trans-tokenization addresses these limitations by explicitly aligning token vocabularies using parallel corpora and statistical alignment, initializing target-language embeddings as weighted combinations of semantically aligned source embeddings (Remy et al., 2024). This approach has been shown to enable stable adaptation of large language models to low-resource languages without catastrophic degradation. In contrast to prior work focusing on cross-lingual transfer for generative models, we adopt transtokenization in a monolingual Arabic

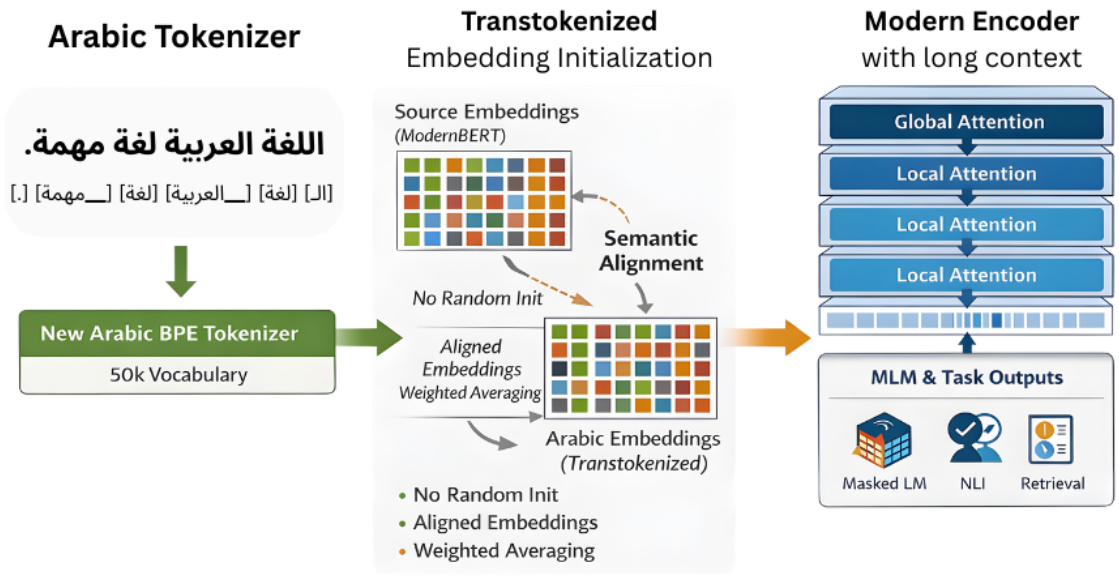

Figure 1: AraModernBERT integrates an Arabic BPE tokenizer with transtokenized embedding initialization and a ModernBERT encoder supporting native long-context modeling up to 8,192 tokens.

setting and demonstrate its critical role in training modern Arabic encoder models.

Finally, long-context modeling and retrieval have received increasing attention as NLP applications move toward document-level understanding. While extended context improves language modeling capacity, prior work shows that naïvely encoding long documents into a single vector often degrades retrieval performance due to representation dilution, motivating multi-vector and late-interaction approaches such as ColBERT (Khattab and Zaharia, 2020; Karpukhin et al., 2020). In Arabic NLP, long-context retrieval remains underexplored, and most systems rely on chunking long documents to fit short-context encoders. Our work contributes empirical evidence to this discussion by analyzing long-context retrieval with a modern Arabic encoder and clarifying when architectural changes beyond context length are required.

3 Methodology

This section describes the design and training of **AraModernBERT**, an Arabic encoder model adapted from the ModernBERT architecture. Our methodology focuses on two central aspects: (i) the transfer of a modernized encoder architecture to Arabic and (ii) the use of transtokenized embedding initialization to enable stable and effective Arabic language modeling. Figure 1 provides an overview of the full pipeline, illustrating how a new Arabic tokenizer is introduced, how its embeddings are initialized via transtokenization, and how the resulting representations are processed by a long-context encoder.

Concretely, given a new Arabic tokenizer and a pretrained source embedding space, transtokenization proceeds by aligning target-language tokens to semantically related source-language tokens using a parallel corpus and statistical alignment. For each Arabic token t, we obtain a set of aligned source tokens $\{s_i\}$ with associated alignment counts $c_{t \to s_i}$. The embedding of t is then initialized as a weighted average of the aligned source embeddings:

$$\mathbf{e}(t) = \sum_i \frac{c_{t \to s_i}}{\sum_j c_{t \to s_j}} \mathbf{e}(s_i), \qquad (1)$$

where $\mathbf{e}(s_i)$ denotes the pretrained embedding of source token s_i, and $c_{t \to s_i}$ is the alignment count between target token t and source token s_i. The normalization ensures that the weights form a probability distribution over aligned source tokens.

For example, the Arabic token اللغة may align to English tokens such as *language* and *linguistic*, and its embedding is initialized as the normalized weighted combination of the corresponding source

embeddings. Tokens without reliable alignments are initialized using predefined fallback mappings (e.g., digits, punctuation, or special symbols). This procedure avoids random initialization while preserving semantic structure in the embedding space.

As shown in Figure 1, transtokenization injects semantically aligned pretrained embeddings into the newly introduced Arabic tokenizer, avoiding the performance degradation typically caused by random embedding initialization. This step is critical for stable masked language model training in Arabic and allows the encoder to fully benefit from the modern architectural features of ModernBERT, including long-context processing.

AraModernBERT is an encoder-only transformer model built on top of the ModernBERT architecture. We retain all core architectural design choices of ModernBERT, which were originally proposed to address efficiency and scalability limitations of classical BERT-style encoders. In particular, AraModernBERT employs a stack of 22 transformer layers with a hidden dimension of 768 and 12 attention heads, resulting in approximately 149 million parameters.

A key feature of the architecture is its *alternating attention mechanism*. Every third layer applies global self-attention, allowing tokens to attend to the entire sequence, while the remaining layers use local self-attention with a sliding window of 128 tokens. This design balances long-range dependency modeling with computational efficiency and enables native processing of long documents.

Context Modeling. AraModernBERT natively supports a maximum sequence length of 8,192 tokens. Long-context capability is enabled through the use of Rotary Positional Embeddings (RoPE), with distinct configuration parameters for global and local attention layers. Specifically, global attention layers use a RoPE theta value of 160,000, while local attention layers use a theta of 10,000. This separation allows the model to maintain positional sensitivity across both short- and long-range interactions.

Importantly, long-context modeling in AraModernBERT is *native* rather than windowed: the full sequence is processed in a single forward pass without truncation or recurrence. This design is particularly well-suited to Arabic-language domains where documents frequently exceed the 512-token limit of traditional encoders.

Arabic Tokenization. Given the morphological richness and orthographic characteristics of Arabic, we train a dedicated Arabic tokenizer rather than reusing multilingual or English-centric tokenizers. The tokenizer is based on byte-pair encoding (BPE) and has a vocabulary size of 50,280 tokens, optimized to capture common Arabic morphemes and word forms while reducing excessive subword fragmentation.

Special tokens follow standard encoder conventions, including dedicated tokens for classification, masking, padding, and separation. This tokenizer serves as the foundation for all pretraining and downstream evaluation.

Transtokenized Embedding Initialization. Replacing a tokenizer in a pretrained model typically requires reinitializing the embedding table, which can lead to severe degradation in performance. To address this issue, AraModernBERT adopts the *transtokenization* strategy for embedding initialization. Transtokenization initializes the embedding vectors of the new Arabic tokenizer using a weighted combination of semantically aligned embeddings from a source model, rather than random initialization. This alignment is derived from cross-lingual token mappings based on translation resources and statistical alignment techniques. By preserving semantic structure in the embedding space, transtokenization enables stable training and effective transfer even when introducing a new tokenizer.

In AraModernBERT, transtokenization is applied to the input embedding layer prior to masked language model training. Our ablation experiments demonstrate that this step is essential for successful Arabic encoder training.

Training Objective and Data. AraModernBERT is trained using the masked language modeling (MLM) objective. During training, 30% of input tokens are masked following standard MLM procedures. Pretraining is conducted on approximately 100 gigabytes of Arabic text drawn from diverse sources, covering a range of domains and writing styles.

Training proceeds in two stages. The model is first trained at shorter sequence lengths to establish stable representations, and subsequently trained with extended sequences up to 8,192 tokens to enable long-context modeling. No task-specific supervision is used during pretraining. Table 1 summarizes the key architectural and training parameters

of AraModernBERT.

Parameter	Value
Architecture	ModernBERT encoder
Hidden size	768
Transformer layers	22
Attention heads	12
Intermediate size	1,152
Vocabulary size	50,280
Maximum context length	8,192
Global attention frequency	Every 3 layers
Local attention window	128 tokens
RoPE theta (global)	160,000
RoPE theta (local)	10,000
Training objective	MLM

Table 1: AraModernBERT configuration and architectural parameters.

4 Experiments and Results

This section presents an empirical evaluation of AraModernBERT across intrinsic language modeling, downstream Arabic natural language understanding, and retrieval. Our experiments are designed to assess three core aspects: (i) the impact of transtokenized embedding initialization, (ii) the effectiveness of native long-context modeling, and (iii) the extent to which the learned representations transfer to downstream Arabic tasks.

4.1 Experimental Setup

We conduct intrinsic evaluations using masked language modeling (MLM) on Arabic Wikipedia. Downstream tasks are evaluated by fine-tuning AraModernBERT with task-specific classification heads on top of the encoder, following standard training protocols. For retrieval, we adopt a dense bi-encoder setup with cosine similarity and in-batch negatives where applicable. All experiments are performed with fixed random seeds and consistent hyperparameter settings to ensure reproducibility.

4.2 Evaluation Metrics

We adopt standard evaluation metrics appropriate for each task. For intrinsic language modeling, we report MLM loss and perplexity, where lower values indicate better modeling performance. For downstream Arabic natural language understanding tasks, we use accuracy for natural language inference and macro-averaged F1 score for classification tasks with class imbalance, including offensive language detection and question-question similarity. For retrieval experiments, we report Recall@k (with $k \in \{1, 5, 10\}$) and Mean Reciprocal Rank (MRR), which measure the ability of the model to rank relevant documents highly. These metrics are widely used in prior work and provide complementary perspectives on model effectiveness across tasks.

4.3 Intrinsic Evaluation: Transtokenization Ablation

To isolate the effect of transtokenized embedding initialization, we compare AraModernBERT against two ablated variants: (i) an embedding reinitialized model, where the tokenizer is kept fixed but the embedding table is randomly reinitialized, and (ii) a fully randomly initialized model with the same architecture.

The results as shown in Table 2, show that transtokenization is critical for Arabic encoder training. Reinitializing the embedding table leads to catastrophic degradation, increasing perplexity by several orders of magnitude. This confirms that embedding initialization plays a central role in stabilizing Arabic language modeling when introducing a new tokenizer.

4.4 Long-Context Language Modeling

We evaluate AraModernBERT under its native 8,192-token context by concatenating Arabic Wikipedia articles into long sequences and computing MLM loss. For comparison, we also report performance at the standard 512-token context length.

Interestingly, as shown in Table 3, MLM loss and perplexity improve at extended context lengths. This indicates that AraModernBERT effectively exploits long-range contextual information rather than suffering from instability or degradation. The model remains memory-efficient, requiring approximately 6.8 GB of GPU memory for 8k-token inference.

4.5 Arabic Natural Language Understanding

We evaluate AraModernBERT on three representative Arabic natural language understanding (NLU) tasks: natural language inference, toxicity detection, and semantic similarity. We use the Arabic subset of XNLI (Conneau et al., 2018), the OSACT4 Offensive Language Detection (OOLD) dataset (Mubarak et al., 2020), and the Mawdoo3 Question Semantic Similarity (MQ2Q) dataset

Model Variant	MLM Loss ↓	Perplexity ↓
AraModernBERT (Transtokenized)	**3.24**	**25.54**
Embedding Re-initialized	11.46	94,372
Fully Random Initialization	10.98	58,962

Table 2: Transtokenization ablation results on Arabic MLM.

Context Length	MLM Loss ↓	Perplexity ↓
512 tokens	3.24	25.54
8,192 tokens	**3.05**	**21.05**

Table 3: Masked language modeling performance at different context lengths.

(Seelawi et al., 2019). For computational consistency, all reported results are obtained on fixed test subsets of 2,000 instances per task. Each task is fine-tuned using a standard classification head on top of the encoder.

Task	Metric	AraModernBERT
XNLI (Arabic)	Accuracy	0.47
OOLD	F1-macro	0.87
MQ2Q	F1-macro	0.96

Table 4: Arabic natural language understanding results.

As shown in Table 4, AraModernBERT demonstrates strong transfer to downstream Arabic NLU tasks, particularly for semantic similarity and offensive language detection. Performance on Arabic XNLI is consistent with prior encoder-based models and reflects the limited size and label noise of available Arabic NLI resources.

4.6 Arabic Retrieval

Short-Text Retrieval. We evaluate short-text semantic retrieval using MQ2Q in a dense bi-encoder setting. Questions are treated as queries and their paired equivalents as relevant documents. AraModernBERT is compared against a representative Arabic encoder baseline, AraBERT-base, under identical training and evaluation conditions.

As shown in Table 5, both models achieve strong retrieval performance. AraBERT slightly outperforms AraModernBERT in this setting, which favors short, lexically similar queries. This result indicates that AraModernBERT remains competitive for short-text semantic retrieval, while its primary advantages lie in representation learning and long-context modeling rather than lexical matching.

4.7 Arabic Named Entity Recognition

Named Entity Recognition (NER) has long been a core task in Arabic NLP, with early systems relying on statistical and rule-based methods tailored to Arabic morphology and orthography (Benajiba et al., 2007). Subsequent work explored the use of cross-lingual resources and multilingual transfer to mitigate data sparsity in Arabic NER (Darwish, 2013; Rahimi et al., 2019). More recent neural approaches have demonstrated strong performance on Arabic NER when sufficient annotated data and appropriate pretraining are available, though performance remains sensitive to domain, noise, and sentence structure (Schneider et al., 2012).

We further evaluate AraModernBERT on Arabic named entity recognition to assess its effectiveness on sequence labeling tasks. Experiments are conducted on multiple Arabic NER benchmarks, including WikiAnn (Arabic) (Rahimi et al., 2019), ANERCorp (Benajiba et al., 2007), and AQMAR (Mohit et al., 2012). All models use a standard token-level classification head and are evaluated using entity-level F1 score, with results averaged over three random seeds.

AraModernBERT achieves its strongest performance on WikiAnn as shown in Table 6, a large-scale and relatively clean NER benchmark with longer average sentence lengths and substantial training data. Performance is more moderate on smaller or noisier datasets such as ANERCorp, AQMAR, and Twitter NER, which include shorter sentences, higher lexical variability, and domain-specific noise. This pattern suggests that AraModernBERT benefits most from settings where richer sentence-level context and larger annotated corpora align with its pretraining regime on long-form, well-structured Arabic text. Similar trends have been observed in prior Arabic NER studies, where encoder-based models trained on clean data exhibit reduced robustness on noisy or informal text (Darwish, 2013; Rahimi et al., 2019).

Across experiments, we find that transtokenization is essential for stable Arabic encoder training and that native long-context modeling improves in-

Model	R@1	R@5	R@10	MRR
AraBERT-base	0.54	0.97	0.99	0.73
AraModernBERT	0.52	0.97	0.99	0.72

Table 5: Short-text retrieval results on MQ2Q.

Dataset	Validation F1	Test F1
WikiAnn (ar)	0.8571	0.8576
ANERCorp	0.8065	0.6827
AQMAR	0.5541	0.5929
Twitter NER	0.5529	0.4919

Table 6: Arabic NER results for AraModernBERT. Scores are entity-level F1 averaged over three seeds.

trinsic language modeling performance. AraModernBERT transfers effectively to downstream Arabic tasks, including natural language understanding, short-text retrieval, and named entity recognition. At the same time, our results highlight that task characteristics and data domain play a central role in determining downstream performance, underscoring the importance of aligning pretraining objectives and data with target applications in Arabic NLP.

5 Discussion

Implications for Arabic Encoder Design. Our experiments demonstrate that tokenizer design and embedding initialization are central to successful Arabic encoder modeling. The transtokenization ablation shows that introducing a new Arabic tokenizer without aligned embedding initialization leads to catastrophic degradation in masked language modeling performance. This finding reinforces the observation that Arabic's morphological richness and lexical sparsity exacerbate tokenizer–embedding mismatches, making careful embedding initialization essential. More broadly, it suggests that future Arabic encoder models should treat tokenizer replacement as a first-class modeling decision rather than a preprocessing detail.

We also show that native long-context modeling can be effectively transferred to Arabic. AraModernBERT remains stable at sequence lengths up to 8,192 tokens and achieves improved intrinsic language modeling performance at extended context lengths. This result is particularly relevant for Arabic domains characterized by long-form text, such as news, legal documents, and encyclopedic content, and supports the feasibility of long-context encoders for Arabic without resorting to windowed or recurrent processing schemes.

Downstream Performance. AraModernBERT transfers effectively to downstream Arabic tasks across both sentence-level classification and sequence labeling. Strong performance on semantic similarity, offensive language detection, and named entity recognition benchmarks demonstrates that gains in intrinsic modeling translate to discriminative settings. In particular, AraModernBERT performs best on larger and cleaner datasets with richer sentence-level context, such as WikiAnn for NER, suggesting alignment between its pretraining regime on long-form Arabic text and downstream data characteristics. More modest results on smaller or noisier datasets, including social media text, are consistent with prior observations for encoder models trained primarily on well-structured corpora.

6 Conclusion

In this work, we introduced **AraModernBERT**, an Arabic adaptation of a modern encoder architecture, and studied the role of tokenizer initialization and long-context modeling for Arabic. Our experiments show that transtokenized embedding initialization is critical for stable Arabic language modeling, leading to substantial improvements in masked language modeling performance. We further demonstrate that AraModernBERT supports native long-context modeling up to 8,192 tokens while remaining computationally efficient. Across downstream evaluations, AraModernBERT transfers effectively to Arabic natural language understanding and sequence labeling tasks, particularly on larger and cleaner datasets with richer sentence-level context. Overall, our findings provide practical guidance for adapting modern encoder architectures to Arabic and other Arabic-script languages.

Limitations

This study has several limitations. While AraModernBERT supports native long-context modeling and demonstrates improved intrinsic performance

at extended sequence lengths, our downstream evaluations focus on tasks that do not explicitly require long-range context at inference time. Evaluating tasks that directly benefit from long-context reasoning, such as document-level information extraction or long-form question answering, represents an important direction for future work. In addition, our experiments are limited to Arabic; although many findings are applicable to other Arabic-script languages, empirical validation on languages such as Persian, Urdu, or Kurdish remains future work. Finally, AraModernBERT is trained on approximately 100 GB of Arabic text, which, while substantial for Arabic, remains modest compared to the scale used for recent English-language encoders.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 7088–7105.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new: How to successfully recycle english gpt-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3118–3135.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse Lexical Semantic Annotation with Supersenses: An Arabic Case Study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258, Jeju Island, Korea. Association for Computational Linguistics.

Haitham Seelawi, Ahmad Mustafa, Hesham Al-Bataineh, Wael Farhan, and Hussein T. Al-Natsheh. 2019. NSURL-2019 task 8: Semantic question similarity in Arabic. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 1–8, Trento, Italy. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Murtadha Ahmed, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.