MarsadLab at BAREC Shared Task 2025: Strict-Track Readability Prediction with Specialized AraBERT Models on BAREC

Shimaa Ibrahim[1], **Md. Rafiul Biswas**[2], **Mabrouka Bessghaier**[1], **Wajdi Zaghouani**[1]
[1]Northwestern University in Qatar
[2]Hamad Bin Khalifa University (HBKU), Qatar
{shimaa.ibrahim,mabrouka.bessghaier,wajdi.zaghouani}@northwestern.edu
mbiswas@hbku.edu.qa

Abstract

The BAREC 2025 Shared Task on Arabic readability targets 19 levels of ordinal prediction at the sentence and document levels under strict training. This paper describes a two stages system that basically starts with BAREC-tuned AraBERT checkpoints and then specializes on the Strict splits with Weighted Kappa Loss (WKL), an objective aligned with Quadratic Weighted Kappa (QWK). A single architecture with inputs specific to each track is utilized for both tracks. On the Strict setting, our best systems reach 0.842/0.841 QWK (public/blind) at the sentence level and 0.828/0.790 QWK at the document level.

1 Introduction

Automatic readability assessment (ARA) estimates how difficult a text is for a target audience. For Arabic, the task is challenging due to morphological richness, orthographic variation, and the coexistence of Modern Standard Arabic (MSA) with regional dialects (Habash, 2010; Cavalli-Sforza et al., 2018). These factors complicate tokenization, feature extraction, and modeling, especially for rare ordinal labels, where small lexical or syntactic differences can shift a sentence between adjacent levels.

The BAREC 2025 Shared Task (Elmadani et al., 2025b) provides a large benchmark with 19 readability levels at the sentence and document levels, spanning multiple domains and genres. Companion resources include a corpus paper (Elmadani et al., 2025a) and detailed annotation guidelines (Habash et al., 2025). We focus on the Strict track, which constrains training to the official data only, resulting in limited data, class imbalance, and closely spaced ordinal labels—conditions that favor pretrained models and ordinal-aware objectives.

Earlier Arabic readability systems relied on manual indicators (e.g., sentence/word length, frequency, morphology) and classical ML, e.g., AARI and OSMAN (Al Tamimi et al., 2014; El-Haj and Rayson, 2016); surveys report that such features under-represent semantics and discourse (Cavalli-Sforza et al., 2018). With Arabic PLMs, performance improved across many tasks (e.g., AraBERT, MARBERT) (Antoun et al., 2020; Abdul-Mageed et al., 2021), but standard fine-tuning with Cross-Entropy (CE) does not align with ordinal evaluation such as Quadratic Weighted Kappa (QWK) (Yannakoudakis et al., 2011).

We propose a two-stage strategy for the Strict track: (i) initialize from BAREC-tuned AraBERT checkpoints, then (ii) fine-tune on the Strict splits with *Weighted Kappa Loss* (WKL), a differentiable surrogate aligned with QWK. We use specific input variants for each track, D3Tok for sentences and Word for documents, and adopt *max-level* aggregation for documents (label = hardest sentence) (Habash et al., 2025). This setup yields strong results at both levels.

2 Background

For education, ARA evaluates reading level to drive text selection, curriculum sequencing, and learner assessment (Vajjala, 2022). Early work relied on manually engineered features such as sentence length, word frequency, and syntactic complexity (Feng et al., 2010; Vajjala, 2022). While effective in controlled settings, such surface features often miss semantic and discourse cues, limiting robustness across genres and languages.

With large pretrained language models (PLMs) such as BERT (Devlin et al., 2019), the field shifted toward holistic fine-tuning with richer contextual representations; recent studies report strong gains for Transformer encoders in readability prediction (Martinc et al., 2021). We defer a focused survey of PLM approaches to Section 3 to avoid redundancy.

For Arabic, readability modeling is particularly challenging due to morphological richness,

Proceedings of The Third Arabic Natural Language Processing Conference, pages 274–279
November 8-9, 2025 ©2025 Association for Computational Linguistics

orthographic variation, and the coexistence of MSA with multiple dialects (Habash, 2010; Nassiri et al., 2023). Concurrent advances in Arabic PLMs—AraBERT (Antoun et al., 2020), ARBERT/MARBERT (Abdul-Mageed et al., 2021), and QARiB (Abdelali et al., 2021)—have delivered strong results across sentiment, dialect identification, and classification benchmarks (Abu Farha and Magdy, 2021); we discuss these in Related Work.

The BAREC resources standardize fine-grained Arabic readability: the shared task overview defines 19 ordinal levels and two evaluation settings ,General and Strict at the sentence and document levels (Elmadani et al., 2025b); the corpus paper details broad coverage for fine-grained labeling (Elmadani et al., 2025a); and the annotation guidelines specify procedures for consistent sentence level judgments (Habash et al., 2025). The **Strict** setting limits training to the official splits, and official evaluation uses *Quadratic Weighted Kappa* (QWK), motivating approaches that leverage pre-trained encoders while aligning optimization with ordinal agreement.

3 Related Work

Early Arabic readability research adapted formulaic, feature-based methods from English, using shallow indicators (e.g., sentence length, word frequency, morphology) and classical ML; systems such as AARI and OSMAN established useful baselines but provide limited coverage of semantics and discourse and transfer poorly across domains (Al Tamimi et al., 2014; El-Haj and Rayson, 2016; Forsyth, 2014; Saddiki et al., 2018; Cavalli-Sforza et al., 2018). With pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), richer contextual representations typically outperform feature-only models on readability prediction (Martinc et al., 2021; Lee et al., 2021). For Arabic NLP, AraBERT, ARBERT/MARBERT, and QARiB advance the state of the art across text classification tasks (Antoun et al., 2020; Abdul-Mageed et al., 2021; Abdelali et al., 2021), motivating PLM-based approaches to Arabic readability.

Readability labels are ordinal; however, optimizing nominal cross-entropy (CE) can misalign with QWK (Yannakoudakis et al., 2011; Martinc et al., 2021). Ordinal aware training includes (i) direct or surrogate optimization of QWK (e.g., WKL) (de la Torre et al., 2018), (ii) regression or threshold based ordinal classification, and (iii) pairwise or ranking objectives, which often reduce large magnitude errors relative to CE.

BAREC standardizes fine-grained Arabic readability with 19 levels at sentence and document scopes and defines Strict,it is data constrained track with settings using only official splits (Elmadani et al., 2025b,a; Habash et al., 2025). Official resources report PLM baselines and fine-grained evaluations; document labels follow the hardest sentence definition (Habash et al., 2025).

Complementary resources provide signals correlated with readability. The SAMER Readability Lexicon and SAMER Simplification Corpus supply leveled lexical cues and aligned simplification pairs, and recent work systematizes strategies for Arabic readability modeling (Al Khalil et al., 2020; Alhafni et al., 2024; Liberato et al., 2024). Orthographic or phonological indicators from large scale diacritized text enable features such as vowelization density and ambiguity reduction (Zaghouani et al., 2016).

Discourse signals arise from punctuation and boundary usage; Arabic punctuation annotation and a punctuated corpus support density of punctuation and restoration models (Zaghouani and Awad, 2016b,a). In learner contexts, correction annotated corpora provide error rate and edit operation statistics that proxy grammaticality and difficulty (Zaghouani et al., 2015). Word-level visualizations further illustrate fine-grained difficulty signals for assisted simplification (Hazim et al., 2022).

Within this landscape, our system starts from BAREC, then tunes PLMs, and continues training with a QWK aligned objective (WKL), targeting Strict track robustness and reduction of large ordinal errors.

4 System Overview

We participate in the **Sentence Strict** and **Document Strict** tracks of BAREC 2025, predicting fine-grained Arabic readability levels (C=19) under constrained training. The setting is challenging due to the large label space, skewed label distribution, and differences between sentence and document level detection.

4.1 Two-Stage Fine Tuning

We adopt a two-stage pipeline. **Stage 1 (warm start):** initialize from public AraBERT-based readability checkpoints released for BAREC (sentence: D3Tok input; document: Word input). These are trained with CE on BAREC and provide domain-driven representations (Antoun et al., 2020; Elmadani et al., 2025a). **Stage 2 (Strict specialization):** fine-tune only on the official Strict splits with WKL, a differentiable surrogate aligned with QWK, penalizing large ordinal errors more than small ones (de la Torre et al., 2018; Yannakoudakis et al., 2011).

Motivation: two-stage CE → WKL. The official metric for BAREC is *Quadratic Weighted Kappa* (QWK), which penalizes larger ordinal mistakes more. We therefore align optimization with evaluation by continuing training using a *Weighted Kappa Loss* (WKL). We use two stages instead of training WKL from scratch because: (i) A CE warm start from a BAREC-tuned checkpoint retains domain and split-specific signals, including tokenization and label priors over 19 levels. (ii) Direct WKL from an untuned PLM exhibited reduced stability on Strict (characterized by class imbalance and narrowly spaced labels), while CE produces a highly accurate classifier that WKL subsequently refines. (iii) Stage 2 emphasizes the mitigation of significant ordinal mistakes that influence QWK with minimal additional procedures. Specifically, upon CE convergence, we reload the checkpoint and transition to WKL with quadratic weights $w_{ij} = \left(\frac{i-j}{K-1}\right)^2$, K=19, lower the learning rate, and apply early stopping on dev QWK.

4.2 Model Architecture

Our model uses a Transformer encoder $\mathcal{E}$ (AraBERT family) with a linear head. Given input x, let $\mathbf{h}_{[\text{CLS}]} = E(x)_{[\text{CLS}]}$. The classifier computes

$$\boldsymbol{\ell} = W\,\mathbf{h}_{[\text{CLS}]} + \mathbf{b}, \quad \mathbf{p} = \text{softmax}(\boldsymbol{\ell}), \qquad (1)$$

where $W \in \mathbb{R}^{C\times d}$, $\mathbf{b} \in \mathbb{R}^{C}$, C=19, and d is the encoder hidden size. As shown in Equation 1, we map [CLS] to logits ℓ then to probabilities p.

4.3 Preprocessing and Optimization

We follow the shared-task input conventions for comparability: D3Tok for sentence-level inputs and Word for document-level inputs (matching the released checkpoints). No external data are used for Strict track. Hyperparameters, includeing learning rate, batch size and warmup, are tuned per track with early stopping on the Strict dev split.

4.4 Document Inference

Document labels are obtained via *max-level pooling* over sentence predictions (document level = level of the hardest sentence), consistent with the task definition (Habash et al., 2025).

4.5 Summary of Differences

In comparison to CE-only baselines using BAREC resources, our system (i) initiates from BAREC-optimized checkpoints, (ii) substitutes CE with WKL in stage 2 to synchronize training with QWK, and (iii) employs track-specific input variations (D3Tok vs Word) in accordance with the sentence/document configuration(Elmadani et al., 2025a).

5 Experimental Setup

We describe the datasets, input variants, model initialization, optimization, and evaluation protocol used in our Strict track sentence and document level experiments.

5.1 Data and Inputs

We use the BAREC 2025 resources, which provide sentence and document level readability annotations across 19 ordered levels (Elmadani et al., 2025b,a; Habash et al., 2025). We follow the official *Strict* splits and do not use external data. For the sentence track, inputs follow the **D3Tok** variant; for the document track, the **Word** variant, matching the released BAREC checkpoints.

5.2 Model Configurations

We adopt a two-stage strategy. **Stage 1** warm-starts from BAREC-tuned AraBERT checkpoints (sentence: D3Tok; document: Word) trained with CE (Antoun et al., 2020; Elmadani et al., 2025a). **Stage 2** specializes in the strict splits using WKL, a differentiable surrogate aligned with QWK, to better reflect ordinal evaluation.

5.3 Training Details

All runs use a single NVIDIA T4 (16 GB). We train with AdamW, initial learning rate 2×10^{-5}, batch size 16, linear decay with warmup ratio 0.1, and early stopping on dev QWK. Each model trains up to 10 epochs; the best dev QWK checkpoint is used for test submission.

5.4 Evaluation Metrics

The official metric is **QWK**,which quantifies agreement while punishing significant ordinal discrepancies. We provide QWK for validation, public test, and blind test partitions; accuracy is assessed only for diagnostic purposes.

Model (Tokenization)	Loss	Val QWK	Public Test QWK	Blind Test QWK
BAREC Official Baseline (Strict leaderboard)	–	–	–	**0.815**
Ours: AraBERTv2 (D3Tok)	WKL	**0.820**	**0.842**	**0.841**

Table 1: Sentence-level *Strict* results (QWK). Baseline taken from the official Strict-track leaderboard

Model (Tokenization)	Loss	Val QWK	Public Test QWK	Blind Test QWK
BAREC Official Baseline (Strict leaderboard)	–	–	–	**0.620**
Ours: AraBERTv2 (Word)	WKL	**0.820**	**0.828**	**0.790**

Table 2: Document-level *Strict* results (QWK). Baseline taken from the official Strict-track leaderboard

6 Results

We demonstrate strict track findings for sentence and document-level tasks, correlate them with corpus-paper baselines when relevant, and analyze observed mistake trends.

Metric. As stated in previous sections, we provide QWK using the official scorer in accordance with the BAREC procedure.

6.1 Sentence-Level (Strict Track)

Table 1 includes the official *Strict*-track baseline from the blind (final) leaderboard (QWK $= 0.815$). Our two-stage CE→WKL approach attains $0.842/0.841$ (public/blind) and improves over this baseline under the same Strict constraints.

6.2 Document-Level (Strict Track)

Table 2 reports our results alongside the official *Strict*-track baseline from the blind (final) leaderboard (QWK $= 0.620$). Our WKL specialization reaches $0.828/0.790$ (public/blind), showing gains on public test and a modest blind drop, suggesting sensitivity to domain shift and to max-level pooling.

Analysis. (1) The implementation of an ordinal-aware objective (WKL) aligns the training process with QWK and is consistent with trends observed in corpus papers, indicating that ordinal objectives demonstrate superior performance compared to CE on development datasets. (2) The sentence-level Strict scores obtained are $0.842/0.841$ for public and blind evaluations, respectively. These scores align with the general range of previous development split results reported on BAREC, even under more stringent training constraints. (3) Document-level blind performance (0.790) lags behind the public benchmark by approximately 0.04, suggesting a sensitivity to shifts in domain or topic as well as to max pooling techniques. Implementing hierarchical document encoders or utilizing calibrated/attention-based aggregation methods may enhance robustness further.
Reproducibility. We will release evaluation scripts, configs, and checkpoints upon acceptance.

7 Conclusion

We examined fine-grained Arabic readability in the *Strict* BAREC 2025 setting by initializing AraBERT from BAREC-tuned checkpoints and fine-tuning using a quadratic, ordinal-aware objective (WKL). An encoder utilizing track-specific inputs (D3Tok for sentences; Word for documents) and max-pooling for document label aggregation achieves **0.842/0.841** QWK (public/blind) at the sentence level and **0.828/0.790** at the document level. Errors are less frequent at higher magnitudes and tend to cluster between neighboring levels. Future research will focus on hierarchical document encoders, advanced aggregation methods beyond max-pooling, and efficient domain/task adaptation under strict constraints.

Limitations

This study is limited to the Strict track and uses only official data and BAREC-tuned checkpoints; generalization to other corpora, domains, or languages is untested. Document labels are obtained by max-pooling sentence predictions, which can be sensitive to outliers and intra-document variation. Compute constraints precluded extensive hyperparameter search or ensembling, and we report single-model runs. Finally, while we optimize an ordinal-aware loss and report QWK, broader evaluation (e.g., MAE, accuracy@±1) and statistical significance across multiple seeds, as well as genre/dialect–level error analysis, are left to future work.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.

Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. 2014. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The SAMER Arabic text simplification corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093, Torino, Italia. ELRA and ICCL.

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, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Current state and future directions. *Procedia Computer Science*, 142:38–49. Arabic Computational Linguistics.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154. Machine Learning and Applications in Artificial Intelligence.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN — a novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255, Portorož, Slovenia. European Language Resources Association (ELRA).

Khalid Elmadani, Nizar Habash, and Hanada Taha. 2025a. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.

Khalid N. Elmadani, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025b. BAREC shared task 2025 on Arabic readability assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Lijun Feng, Michael Elhadad, and Matt Huenerfauth. 2010. Automatic readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–62.

Jonathan Forsyth. 2014. Automatic readability prediction for modern standard arabic. Master's thesis, Brigham Young University, Provo, UT.

Nizar Habash. 2010. *Introduction to Arabic natural language processing*, 1 edition, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.

Nizar Habash, Hanada Taha-Thomure, Khalid Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. Guidelines for fine-grained sentence-level Arabic readability annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 359–376, Vienna, Austria. Association for Computational Linguistics.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juan Liberato, Bashar Alhafni, Muhamed Khalil, and Nizar Habash. 2024. Strategies for Arabic readability modeling. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 55–66, Bangkok, Thailand. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2023. Approaches, methods, and resources for assessing the readability of arabic texts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil. 2018. Feature optimization for predicting readability of Arabic L1 and L2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29, Melbourne, Australia. Association for Computational Linguistics.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Wajdi Zaghouani and Dana Awad. 2016a. Building an arabic punctuated corpus. 2016(1):SSHAPP3148.

Wajdi Zaghouani and Dana Awad. 2016b. Toward an arabic punctuated corpus: Annotation guidelines and evaluation. In *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, volume 22.

Wajdi Zaghouani, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, and Kemal Oflazer. 2016. Guidelines and framework for a large scale Arabic diacritized corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3637–3643, Portorož, Slovenia. European Language Resources Association (ELRA).

Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider, and Kemal Oflazer. 2015. Correction annotation for non-native Arabic texts: Guidelines and corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 129–139, Denver, Colorado, USA. Association for Computational Linguistics.