ZAI at BAREC Shared Task 2025: AraBERT CORAL for Fine Grained Arabic Readability

Ahmad M. Nazzal
ZAI Arabic Language Research Center / Zayed University

Email: Ahmad.Nazzal@zu.ac.ae

Abstract

Readability assessment is essential for effective communication of scientific and medical content in Arabic. We present a system for the BAREC 2025 Shared Task Arabic Readability Assessment. The system fine-tunes AraBERTv2 with a CORAL ordinal head, applies AraBERT-specific preprocessing, and selects checkpoints using Quadratic-Weighted Kappa (QWK) with early stopping. Our model achieves a QWK of 85.5 on the Sentence Blind Test, demonstrating its effectiveness for automatic Arabic readability prediction.

1 Introduction

Automatic readability assessment supports education, accessibility, and editorial workflows by estimating how difficult a text is for a target audience (Hazim et al., 2022; Liberato et al., 2024). For Arabic, this task is especially challenging: the language is morphologically rich and it exhibits diglossia between Modern Standard Arabic and regional dialects, which complicates lexical and morpho-syntactic cues used by models (Asadi and Abu-Rabia, 2019; Ferguson, 1959; Saiegh-Haddad and Ghawi-Dakwar, 2017; Taha and Saiegh-Haddad, 2016). The BAREC 2025 shared task addresses these challenges with a large, fine-grained sentence-level benchmark annotated into 19 ordered levels, accompanied by clear guidelines and an evaluation based on quadratic-weighted kappa (QWK) (Al Khalil et al., 2020; Elmadani et al., 2025a,b; Habash et al., 2025). We participate in the Sentence-level (Strict) track. We present a compact, reproducible system: AraBERTv2 (Antoun et al., 2020) with a rank-consistent ordinal regression (CORAL) head (Aicher et al., 2022; Cao et al., 2020). Training uses early stopping with QWK-based model selection; inference applies a single development-tuned threshold to convert cumulative probabilities into one of the 19 levels (no temperature scaling). This simple architecture achieves strong performance on the Blind Test.

2 Background

Early work estimated readability using surface proxies such as sentence/word length and syllabification, yielding indices like Flesch Reading Ease and Dale–Chall (Dale and Chall, 1948; Flesch, 1948). Contemporary approaches treat readability as supervised prediction over lexical, syntactic, and distributional features, increasingly with pretrained language models. Fine-grained readability labels are ordered; modeling them as nominal classes discards rank information. Ordinal regression methods—especially CORAL, which learns $K-1$ binary thresholds for events $y > k$—enforce label order and are often preferable to softmax for such targets. As resources, the BAREC benchmark provides a large sentence-level corpus with 19 readability levels, detailed annotation guidelines, and official splits (*Open Dev*, *Open Test*, *Blind Test*) tailored for shared-task evaluation (Elmadani et al., 2025a,b; Habash et al., 2025). Related Arabic resources such as SAMER target text simplification rather than graded readability but reflect a broader interest in accessibility for Arabic texts (Alhafni et al., 2024; Al Khalil et al., 2020). Given these factors—Arabic's linguistic properties, the ordinal nature of labels, and QWK as the official metric—pretrained Arabic encoders such as AraBERT offer a natural foundation for readability systems; we therefore build on AraBERTv2 and an ordinal head in the methods that follow (Antoun et al., 2020).

3 System Overview

3.1 Encoder and Preprocessing

We Fine-tune aubmindlab/bert-base-arabertv2. We start by normalizing/segmented sentences with the AraBERTPreprocessor and tokenized with fixed max length = 128 (no dynamic padding).

Proceedings of The Third Arabic Natural Language Processing Conference, pages 266–268

3.2 Ordinal Head (CORAL)

Let $K = 19$ and labels $y \in \{0, \ldots, K-1\}$ (with labels shifted by -1 during training). The ordinal head outputs logits z_k for $k = 1, \ldots, K-1$; after applying the sigmoid function, $\sigma(z_k) \approx P(y > k)$. The training targets are defined as cumulative indicators

$$t_k = \begin{cases} 1, & y > k \\ 0, & \text{otherwise} \end{cases}$$

The loss function is the binary cross-entropy with logits, summed over all thresholds:

$$L = \sum_{k=1}^{K-1} \text{BCEWithLogits}(z_k, t_k),$$

thereby enforcing consistent ordering of the predicted categories (Cao et al., 2020).

3.3 Training and Selection

Optimization uses AdamW, a linear schedule with warmup, gradient clipping, label smoothing $= 0.0$ (smoothing hurt this fine-grained ordinal task), and early stopping (patience $= 2$). We set `metric_for_best_model = eval_qwk` and `greater_is_better = True` so the saved model maximizes QWK.

3.4 Inference (Single Threshold Only)

We convert the $(K-1)$ probabilities to a level by counting how many exceed a single threshold t, tuned on the dev set to maximize QWK. No temperature scaling or additional calibration is used in the final system.

4 Experimental Setup

We use the organizers' *Open Train*, *Open Dev*, *Open Test*, and *Blind Test* (sentence track). Splits are unchanged. Key hyperparameters: Encoder: AraBERTv2; max length $= 128$. Optimizer: AdamW (lr $= 2 \times 10^{-5}$), weight decay $= 0.01$; linear decay; warmup $= 6$; max-grad-norm $= 1.0$. Batching: train batch size $= 16$, eval batch size $= 32$; gradient accumulation $= 2$. Regularization: label smoothing $= 0.0$; early stopping patience $= 2$. Precision: FP16 on T4 (BF16 if available). Selection: best checkpoint by eval_qwk; threshold t tuned on dev. Implementation: Transformers 4.54.0; Datasets ≥ 2.18.

Data	QWK	Acc. (%)	Acc. ±1 (%)	Dist.	Acc. 7	Acc. 5	Acc. 3
Open Dev.	66.6	41.5	55.8	1.6	51.9	57.4	64.9
Open Test	72.9	46.6	61.2	1.4	55.8	61.0	69.3
Blind Test	85.5	35.6	74.2	1.0	64.4	69.3	75.8

Table 1: Model performance across datasets. QWK = quadratic weighted kappa.

5 Results

On the Open Dev set, the model reached a QWK of 66.6; performance improved on the Open Test set (72.9) and peaked on the Blind Test set (85.5). Accuracy was moderate overall (35–47%), but accuracy within one level was substantially higher (56–74%), indicating the model captures ordinal trends even if exact prediction is difficult. The distribution score decreased from 1.6 on Dev to 1.0 on Blind, suggesting better calibration on held-out data. As expected, replacing CORAL with a nominal softmax head reduced QWK, confirming the benefit of enforcing label order. Label smoothing and temperature scaling both impaired dev QWK, so the final system uses neither. No explicit error analysis was conducted. Results are summarized in Table 1.

6 Conclusion

We presented a compact system for fine-grained Arabic readability in the BAREC 2025 shared task. The method combines AraBERTv2 with a CORAL ordinal head, trains with QWK-based model selection and early stopping and predicts with a single development-tuned threshold. Without external data or complex ensembling, the system achieves QWK = 85.5 on the Sentence Blind Test. The results support two takeaways: (i) respecting label order via an ordinal head is effective for 19-level readability; and (ii) aligning selection and post-processing with the official metric (QWK) is a simple, high-leverage choice.

Limitations

We rely solely on the shared task splits, no external corpora or augmentation. Domain transfer beyond BAREC is untested. A single encoder is used; larger backbones or multilingual pretraining were not explored due to time/compute. No document-level context or explicit linguistic features (e.g., morphological complexity, type–token ratio) are used. We focus on core choices (ordinal vs. nominal, smoothing, temperature). Future work includes error analysis, and exploring alternative or-

dinal objectives (e.g., CORN) and document-level context.

Ethics Statement

The author declares an affiliation with an institution that contributed to the preparation of the shared task. None of the organizers contributed to the conception, development, or evaluation of our systems. All information and resources used were based exclusively on resources publicly released to all participants, without any form of privileged access or guidance.

References

Annalena Aicher, Alisa Gazizullina, Aleksei Gusev, Yuri Matveev, and Wolfgang Minker. 2022. Towards speech-only opinion-level sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2000–2006, Marseille, France. European Language Resources Association.

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.

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.

Ibrahim A. Asadi and Salim Abu-Rabia. 2019. The impact of the position of phonemes and lexical status on phonological awareness in the diglossic arabic language. *Journal of Psycholinguistic Research*, 48(5):1051–1062.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

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

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025b. 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.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Rudolf. Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Nizar Habash, Hanada Taha-Thomure, Khalid N. 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.

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.

Elinor Saiegh-Haddad and Ola Ghawi-Dakwar. 2017. Impact of diglossia on word and non-word repetition among language impaired and typically developing arabic native speaking children. *Frontiers in Psychology*, 8:2010.

Haitham Taha and Elinor Saiegh-Haddad. 2016. The role of phonological versus morphological skills in the development of arabic spelling: An intervention study. *Journal of Psycholinguistic Research*, 45(3):507–535.