mucAI at BAREC Shared Task 2025: Towards Uncertainty Aware Arabic Readability Assessment

Ahmed Abdou
Independent Researcher. Munich, Germany
ahmedabdou1789@gmail.com

Abstract

We present a simple, model-agnostic post-processing technique for fine-grained Arabic readability classification in the BAREC 2025 Shared Task (19 ordinal levels). Our method applies conformal prediction to generate prediction sets with coverage guarantees, then computes weighted averages using softmax-renormalized probabilities over the conformal sets. This uncertainty-aware decoding improves Quadratic Weighted Kappa (QWK) by reducing high-penalty misclassifications to nearer levels. Our approach shows consistent QWK improvements of 1-3 points across different base models. In the strict track, our submission achieves QWK scores of 84.9%(test) and 85.7% (blind test) for sentence level, and 73.3% for document level. For Arabic educational assessment, this enables human reviewers to focus on a handful of plausible levels, combining statistical guarantees with practical usability.

1 Introduction

Automatic readability assessment estimates how difficult a text will be for a target audience, a task essential for the design and advancement of pedagogically oriented NLP applications(Collins-Thompson and Callan, 2004; Xia et al., 2016). In Arabic, this problem is particularly challenging due to morphological richness, and orthographic variation (Liberato et al., 2024; Benajiba and Rosso, 2008). Recent work has advanced Arabic readability assessment through modeling and datasets (Saddiki et al., 2018; Alhafni et al., 2024; Elmadani et al., 2025a; Habash et al., 2025). Most recently, the BAREC corpus (Elmadani et al., 2025a) which offers 19 fine-grained levels. Nevertheless, even state-of-the-art models like AraBERT-v2 (Antoun et al., 2020) remain prone to large-gap misclassifications and offer no principled means of quantifying prediction uncertainty. We address this by integrating conformal prediction (Vovk et al., 2005) to produce statistically valid prediction sets and uncertainty-guided final predictions, reducing high-penalty errors and enabling compact, interpretable outputs for human-in-the-loop educational use. On the BAREC 2025 Shared Task, our method consistently improves QWK across base models, reaching 84.9% on the test set and 85.7% on the blind test at the sentence level, and 73.3% on the blind test at the document level. Beyond leaderboard improvements, our method provides interpretable prediction sets and uncertainty estimates that enable more reliable readability assessment. Our implementation is open-sourced for reproducibility[1].

2 Background

2.1 Task and Data

The BAREC Shared Task 2025 (Elmadani et al., 2025b) targets fine-grained Arabic readability assessment across 19 ordered levels. The task builds on the BAREC corpus (Elmadani et al., 2025a), a manually annotated dataset containing over 69,000 sentences and more than one million words. The corpus provides mappings to multiple granularities (3, 5, and 7 readability levels); for detailed annotation guidelines, we refer readers to (Habash et al., 2025). We participated in both sentence-level and document-level variants of the strict track, where participants are restricted to using only the BAREC corpus for training. In the document-level task, a document's overall readability level is determined by its most difficult sentence. Given the ordinal nature of readability levels, the main evaluation metric is Quadratic Weighted Cohen's Kappa (QWK), which penalizes larger misclassifications more heavily (Cohen, 1968). This reflects the educational goal of avoiding assignments far from a student's level. We also report exact accuracy, ad-

[1] https://github.com/AhmedAbdel-Aal/mucAI-at-BAREC_2025

Proceedings of The Third Arabic Natural Language Processing Conference, pages 312–319
November 8-9, 2025

jacent accuracy (±1 of true label), Mean Absolute Error (MAE), and coarse-grained variants Acc7, Acc5, and Acc3, which collapse the 19 levels into 7, 5, and 3 bins. The shared task provides standard splits: training (54.8k), development (7.3k), test (7.3k), and blind test (3.4k), with the first three publicly available[2].

2.2 Conformal Prediction

Conformal Prediction (CP) (Vovk et al., 2005; Papadopoulos et al., 2002) is a model-agnostic method that converts single predictions into prediction sets with statistical guarantees. Rather than predicting "this text is Level 9", CP produces "this text is likely Level 7, 8, 9, 10, or 11". Given a target miscoverage rate α, CP guarantees that the true label appears in the prediction set with probability at least $1 - \alpha$:

$$P\big(Y \in C(X)\big) \geq 1 - \alpha \tag{1}$$

where $C(X)$ is the predicted set for input X and Y is the true label. The method works by using a calibration set, data not seen during training, to learn how "unusual" different labels are for given inputs. This unusualness is captured by a nonconformity score $s(x, y)$: higher scores mean label y is less plausible for input x (more in appendix A.1.). CP then sets a threshold $\hat{\tau}$ which is chosen as the $(1 - \alpha)(n + 1)$-quantile of these scores in the calibration set, ensuring the coverage guarantee. For any new input x, the prediction set includes all labels below this threshold:

$$C(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{\tau}\} \tag{2}$$

3 Method

We use AraBERT-v2 (Antoun et al., 2020) as the backbone, following the strongest BAREC baselines (Elmadani et al., 2025a). The original benchmark reports four preprocessing pipelines based on CAMeL tools (Obeid et al., 2020) (Word, Lex, D3Lex, D3Tok) but we could not run the CAMeL D3 analyzer in our environment. Because BAREC releases the dev/test sentences already preprocessed with these pipelines, we include them for comparison. For the blind split, however, only raw text is provided; we therefore adopt AraBERT's recommended Farasa segmentation (Abdelali et al., 2016). For training objectives, we replicate the benchmark baselines: Cross-Entropy (CE) and Earth Mover's Distance (EMD) (Hou et al., 2017), and an ordinal Regression variant. Our addition is a Focal-loss objective (Lin et al., 2017) tailored to the long-tailed 19-level label distribution; we report it alongside the baselines and simple ensembles: probability averaging, and majority voting.

Our post-processing approach combines conformal prediction with expected value decoding. We first generate prediction sets with coverage guarantees, then produce final predictions by averaging within these sets. We apply CP only to the probabilistic classifiers (CE/EMD/Focal); the Regression head is reported as point predictions only.

Prerequisites and Notation. Let $\mathcal{Y} = \{1, ..., 19\}$ denote the ordered labels. A trained classifier produces posterior probabilities $p(y \mid x)$ for input x. For any x, we build form a conformal prediction set $C(x) \subseteq \mathcal{Y}$ and then decode to a single label.

Calibration and Tuning Protocol. We split the official development set into two stratified halves: a calibration split (*dev-cal*) for learning conformal thresholds, and a tuning split (*dev-tune*) for hyperparameter selection and evaluation. See the split details in Table 5 in Appendix A.5.

Set Construction. We evaluate three standard nonconformity score functions for multiclass conformal prediction: naïve (inverse-probability), APS (Adaptive Prediction Sets) (Romano et al., 2020), and RAPS (Regularized APS) (Angelopoulos et al., 2020).

Renormalization within the set. We first renormalize probabilities within the conformal set

$$p_C(y \mid x) = \frac{p(y \mid x)}{\sum_{j \in C(x)} p(j \mid x)} \quad \text{for } y \in C(x).$$

We then predict the rounded posterior mean

$$\hat{y}(x) = \text{round}\left(\sum_{y \in C(x)} y \, p_C(y \mid x) \right).$$

The choice of weighted mean is motivated by its role as the Bayes-optimal point estimator under quadratic loss. While this is not strictly optimal for our discrete classification setting, we employ it as a computationally simple heuristic that aligns with the quadratic penalty structure of the primary evaluation metric (QWK). For the document-level

[2] BAREC corpus shared-task-2025

track, we applied our best-performing sentence-level model to all sentences in a document and assigned the document's readability as the maximum predicted level across its sentences, following the shared task definition. We report the full experimental setup in appendix A.2.

4 Results

Dev/test results demonstrate that clitic-aware preprocessing substantially improves performance: Farasa and D3Tok consistently outperform word-level and lexical baselines, with Farasa achieving the best QWK scores under CE, EMD, and regression losses, and on par under Focal loss. Given Farasa's consistent performance across dev/test splits and its availability as the only accessible preprocessor for blind evaluation, we standardize on Farasa preprocessing for all subsequent experiments (full results in Appendix A.4).

Table 1 reports sentence-level results on the BAREC 2025 test set. +CP improves QWK over each baseline while reducing exact Acc, and increases ±1Acc. The strongest single model is Focal+CP (QWK 84.4; +2.6 over Focal); CE+CP and EMD+CP gain +1.6 and +1.1 QWK, respectively. The Avg and Most Common ensembles also improve QWK (to 84.9 and 84.6) and reduce Dist (down to 1.01). To quantify headroom if a user could reliably choose from the CP set, we add a non-deployable Oracle: it selects the gold label whenever it lies in the CP set, otherwise falls back to Focal+CP. This upper bound reaches QWK 95.3 and Acc 94.8, closely tracking the target coverage (α=0.10), and illustrates the potential of human-in-the-loop use of CP sets. Results on the blind test set (Table 2) validate the robustness of our approach. The ensemble averaging method achieves the highest performance at 85.7 QWK, while individual CP-enhanced models reach competitive scores of 84.3 (CE), 84.6 (EMD), and 85.3 (Focal). The regression baseline achieves 85.41 QWK, demonstrating strong performance of the regression formulation without post-processing. The consistent pattern of QWK improvements across different loss functions and evaluation sets demonstrates the generalizability of our conformal prediction approach.

5 Discussion

We analyze our conformal prediction approach with the focal loss model and APS at $\alpha = 0.1$, the best-performing setting on the dev-tune split.

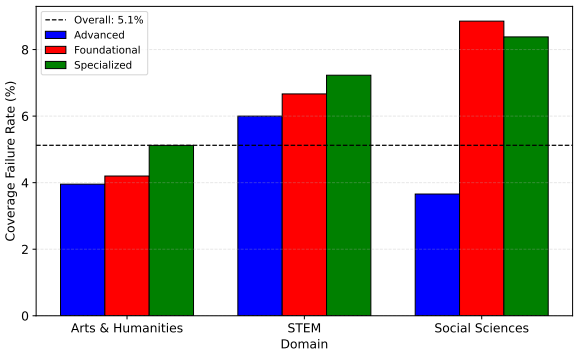

Figure 1: Coverage failure rates by domain and text class. Each domain shows three grouped bars representing Advanced, Foundational, and Specialized text classes. The dashed line shows the overall failure rate (5.12%).

The analysis highlights two aspects: (1) coverage reliability and failure patterns, (2) error redistribution underlying improvements in ordinal metrics.

5.1 CP Coverage Analysis

Using $\alpha = 0.1$ targeting 90% coverage, we report 94.88% empirical coverage with an average set size of 5 levels, a substantial reduction from the full 19-class space. This means that in nearly 95% of Arabic texts, the correct readability level appears in a compact, interpretable set. The remaining 5.12% coverage failures show systematic domain variation: 4.3% for Arts & Humanities (70/1,625), 6.1% for STEM (10/163), and 7.1% for Social Sciences (38/535). We define failure rate as the proportion of cases where the true label falls outside the conformal prediction set. Figure 1 reveals that failures are not uniformly distributed across text types. Social Sciences exhibits the highest rates, particularly for Foundational and Specialized texts (8-9% failure rates), while Arts & Humanities remains close to the overall rate. STEM shows elevated failure rates (6-7%) across all text classes. This variation suggests that domain-adaptive calibration strategies could improve coverage reliability for challenging text types. Additional coverage diagnostics are provided in Appendix A.3.

5.2 Why QWK improves despite lower exact accuracy

QWK increases because many large errors shrink while only a smaller set of perfect predictions become near misses. On the dev–tune split, CP turned 362 perfect predictions into errors (15.6%), and 86.7% of these new errors were only ± 1 level.

Model Variant	QWK	Acc^{19}	±1 Acc^{19}	Dist	Acc^{3}	Acc^{5}	Acc^{7}
CE (Baseline)	82.6	**55.5**	71.6	1.04	79.8	71.4	**65.4**
CE + CP	84.3	50.3	72.9	1.03	**80.1**	70.1	63.8
EMD (Baseline)	82.8	54.4	71.4	1.04	79.7	**71.5**	64.6
EMD + CP	83.9	49.4	73.4	1.04	79.7	70.4	63.3
Focal (Baseline)	81.8	55.4	71.7	1.07	79.7	71.4	65.3
Focal + CP	84.4	42.7	**74.5**	1.08	78.0	67.9	61.0
Regression (Baseline)	83.8	42.0	73.2	1.12	78.0	67.3	59.8
Average	**84.9**	47.3	74.0	1.03	79.8	69.6	63.0
Most Common	84.6	49.6	74.4	**1.01**	**80.1**	70.9	64.4
Oracle Decoder	95.3	94.8	95.3	0.20	96.4	95.6	95.3

Table 1: BAREC test, sentence-level. "Baseline" = fine-tuned point decoder. "+CP" = conformal prediction (α=0.10) with our QWK-aligned mean-in-set decoder; applied to CE/EMD/Focal only (Regression is point-only). "Oracle" = upper bound that selects the gold label if it lies in the CP set; otherwise falls back to Focal+CP. All results use Farasa preprocessing.

Model Variant	QWK
CE (Baseline)	82.6
CE + CP	84.3
EMD (Baseline)	-
EMD + CP	84.6
Focal (Baseline)	-
Focal + CP	85.3
Regression (Baseline)	85.4
Average	**85.7**
Most Common	84.8
Document-level (Max over sentences)	73.3

Table 2: Blind test set QWK results. Missing baseline values (–) indicate models not submitted without CP enhancement. Document-level results use the maximum predicted sentence-level difficulty per document.

At the same time, 397 originally incorrect predictions improved (17.1%): 80.6% shrank by 1 level, 14.7% by 2, 3.1% by 3, and 1.6% by 4. Since QWK penalizes errors by the *squared* distance, shrinking many large mistakes yields big gains (e.g., reducing a 4-level error to 1 cuts the penalty from 16 to 1).

6 Conclusions and Future Work

We presented a simple, model-agnostic post-processing method for Arabic readability assessment that combines conformal prediction with expected value decoding. Applied to the BAREC Shared Task 2025, our approach achieved consistent QWK gains of 1-3 points across multiple base models. In the strict track, our submission achieves QWK scores of 84.9% (test) and 85.7% (blind test) for sentence level, and 73.3% for document level. Beyond leaderboard gains, the method produces compact prediction sets with statistical coverage guarantees, offering both improved accuracy and interpretable outputs for human-in-the-loop use.

Future work could extend this approach in several ways. Mondrian conformal prediction could calibrate separately for different text types or complexity ranges, potentially reducing coverage failures in difficult cases. Multi-granularity training using the BAREC mappings (3-, 5-, and 7-level schemes) may improve generalization across difficulty levels. Finally, rule-based or heuristic decoding strategies informed by the official annotation guidelines (Habash et al., 2025) could refine label selection from CP sets by leveraging linguistic cues and common annotation patterns.

7 Limitations

While our approach improves QWK and reduces high-penalty errors, several limitations remain. Most error reductions occur within medium difficulty ranges, leaving large-gap errors at higher levels (e.g., 15–19) largely unresolved. The effectiveness of our approach depends on the base model's calibration: overconfident but incorrect probability estimates can lead to suboptimal conformal sets, and renormalization may not fully correct such biases. Finally, our CP implementation yields slightly conservative coverage (94% vs. 90% target), suggesting room for tighter calibration or adaptive thresholding.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

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.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

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.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Kevyn Collins-Thompson and Jamie Callan. 2004. Information retrieval for language tutoring: An overview of the reap project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545.

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.

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.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2017. Squared earth movers distance loss for training deep neural networks on ordered-classes. In *NIPS workshop*, volume 5.

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference*, pages 7022–7032.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591.

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.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

A Appendix A

A.1 Nonconformity Scores

In conformal prediction, a *nonconformity score* $s(x, y)$ quantifies how atypical a candidate label y is for an instance x given the model's output distribution $p(y \mid x)$. We evaluate three standard multiclass scoring functions:

Naïve (Inverse Probability). The simplest approach uses the complement of the predicted probability:

$$s_{\text{naive}}(x, y) = 1 - p(y \mid x) \tag{3}$$

This yields smaller scores for high-probability labels, producing larger prediction sets for low-confidence predictions.

Adaptive Prediction Sets (APS) (Romano et al., 2020). Let $\pi_1, \pi_2, \ldots, \pi_K$ denote the classes sorted in descending order of their probabilities $p(\pi_1 \mid x) \geq p(\pi_2 \mid x) \geq \cdots \geq p(\pi_K \mid x)$. For a given label y, let $r(y)$ be its rank in this sorted order. The APS score is the cumulative probability mass up to and including label y:

$$s_{\text{aps}}(x, y) = \sum_{j=1}^{r(y)} p(\pi_j \mid x) \tag{4}$$

Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos et al., 2020). RAPS extends APS by adding a linear rank-based penalty:

$$s_{\text{raps}}(x, y) = \sum_{j=1}^{r(y)} p(\pi_j \mid x) + \lambda \cdot r(y) \tag{5}$$

where $\lambda \geq 0$ is the regularization parameter controlling the size-coverage trade-off. In this work, we set $\lambda = 0.01$.

A.2 Experimental Setup

All experiments were conducted on a single NVIDIA A100 GPU using Google Colab Pro. Training was performed for 6 epochs with a batch size of 64, a learning rate of 5×10^{-5}, and the Adam optimizer. The best checkpoint was selected based on development set performance measured by Quadratic Weighted Kappa (QWK).

A.3 CP Coverage Plots

To better understand the behavior of our conformal prediction variants, we provide supplementary plots analyzing performance, coverage calibration, and set size trends across different miscoverage rates α. In Figure 2, we show the relationship between miscoverage rate α and Quadratic Weighted Kappa (QWK) for three conformal prediction methods on the dev-tune set. APS and RAPS maintain stable QWK across all α values, consistently outperforming the baseline. The naïve method degrades sharply beyond $\alpha > 0.2$, indicating poor robustness when allowing larger miscoverage. In Figure 3, we plot the actual coverage against the target coverage for Naïve, APS, and RAPS methods. All methods achieve coverage above the target across the range, indicating slight conservativeness. This effect is most pronounced for APS, which consistently overshoots the target coverage. Such conservative calibration ensures statistical validity but may produce larger prediction sets than necessary, potentially impacting their interpretability. Finally, figure 4 shows the relationship between the miscoverage rate α and the average prediction set size for the three nonconformity scoring methods. For α, APS and RAPS yield larger sets than the naïve method, with APS producing the widest sets.

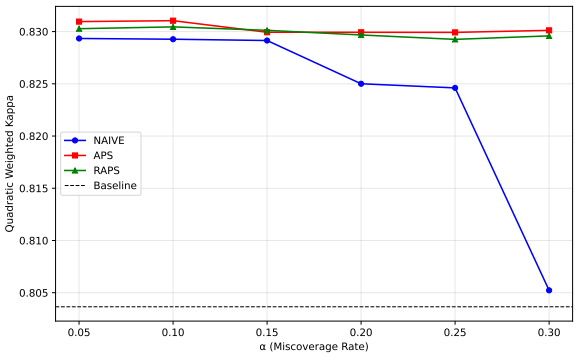

Figure 2: Quadratic Weighted Kappa performance vs. miscoverage rate (α) for three conformal prediction scoring methods on the dev-tune split. The dashed line represents baseline performance without conformal prediction.

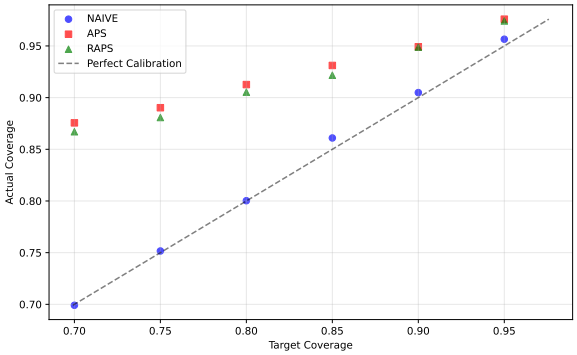

Figure 3: Coverage calibration quality showing actual vs. target coverage rates. The dashed line represents perfect calibration where actual coverage equals target coverage.

A.4 Preprocessing & Loss Ablations

A.5 Dev Data Split

Loss	Input	QWK	Acc^{19}	±1 Acc^{19}	Dist
CE	Word	77.6	53.4	68.2	1.24
CE	Lex	76.4	49.0	66.1	1.32
CE	D3Lex	79.8	53.0	68.3	1.19
CE	D3Tok	81.4	53.3	**70.9**	1.14
CE	Farasa	**80.2**	**55.5**	70.6	**1.13**
EMD	Word	78.2	52.0	67.3	1.24
EMD	Lex	79.5	48.8	66.8	1.24
EMD	D3Lex	80.4	52.2	68.3	1.18
EMD	D3Tok	81.2	53.1	69.8	1.13
EMD	Farasa	**81.4**	**54.8**	**71.0**	**1.10**
Regression	Word	79.3	38.5	69.4	1.30
Regression	Lex	80.9	35.8	69.2	1.31
Regression	D3Lex	82.3	38.7	70.7	1.26
Regression	D3Tok	82.4	40.7	71.5	1.20
Regression	Farasa	**82.9**	**43.3**	**72.5**	**1.15**
Focal	Word	77.6	52.6	67.6	1.25
Focal	Lex	77.9	49.4	67.0	1.27
Focal	D3Lex	80.0	53.4	69.1	1.18
Focal	D3Tok	**80.5**	56.0	**71.1**	**1.12**
Focal	Farasa	80.4	**56.1**	71.0	**1.12**

Table 3: AraBERTv2 results on the BAREC Development set across different loss functions and input representations.

Loss	Input	QWK	Acc^{19}	±1 Acc^{19}	Dist
CE	Word	79.2	54.0	68.6	1.17
CE	Lex	78.4	49.7	66.9	1.23
CE	D3Lex	80.6	53.2	68.1	1.14
CE	D3Tok	81.9	52.8	70.9	1.10
CE	Farasa	**82.6**	**55.5**	**71.6**	**1.04**
EMD	Word	80.7	53.3	68.9	1.13
EMD	Lex	80.6	49.6	67.0	1.18
EMD	D3Lex	81.3	53.3	69.6	1.11
EMD	D3Tok	81.7	52.7	69.3	1.10
EMD	Farasa	**82.8**	**54.5**	**71.4**	**1.04**
Regression	Word	81.4	38.8	70.4	1.23
Regression	Lex	81.4	35.5	70.1	1.26
Regression	D3Lex	82.8	39.2	70.9	1.18
Regression	D3Tok	83.1	40.7	72.2	1.15
Regression	Farasa	**83.8**	**42.0**	**73.2**	**1.11**
Focal	Word	79.9	53.9	69.4	1.14
Focal	Lex	79.5	50.6	67.7	1.19
Focal	D3Lex	80.9	53.1	69.6	1.13
Focal	D3Tok	**82.2**	55.2	71.2	**1.06**
Focal	Farasa	81.8	**55.4**	**71.7**	1.07

Table 4: AraBERTv2 results on the BAREC Test set across different loss functions and input representations.

Class	Original	%	Dev-Cal	Dev-Tune	Split Ratio
1	44	0.6	32	12	73:27
2	68	0.9	49	19	72:28
3	182	2.5	126	56	69:31
4	78	1.1	55	23	71:29
5	417	5.7	284	133	68:32
6	189	2.6	130	59	69:31
7	701	9.6	476	225	68:32
8	613	8.4	417	196	68:32
9	236	3.2	162	74	69:31
10	1012	13.8	686	326	68:32
11	409	5.6	279	130	68:32
12	1491	20.4	1010	481	68:32
13	349	4.8	239	110	68:32
14	1072	14.7	727	345	68:32
15	258	3.5	177	81	69:31
16	114	1.6	80	34	70:30
17	49	0.7	36	13	73:27
18	13	0.2	10	3	77:23
19	15	0.2	12	3	80:20
Total	**7310**	**100.0**	**4981**	**2329**	**68:32**

Table 5: Development set stratified split into calibration (Dev-Cal) and tuning (Dev-Tune) subsets.

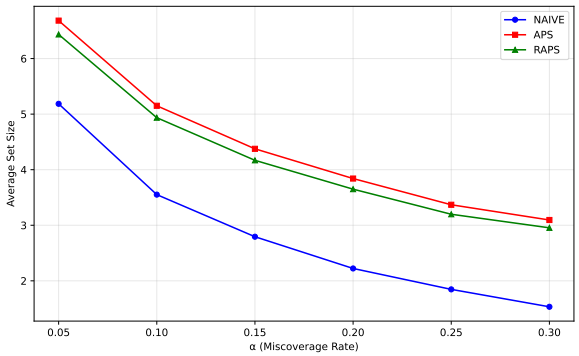

Figure 4: Average prediction set sizes across miscoverage rate (α) for the three conformal prediction scoring methods.