AMAR at BAREC Shared Task 2025: Arabic Meta-learner for Assessing Readability

Mostafa Saeed,[1] Rana Waly,[2] Abdelaziz Ashraf Hussein[3]
[1]New York University Abu Dhabi, UAE
[2]Digital Egypt for Investment Co., Cairo, Egypt
[3]Graduate School of Science and Engineering, Ozyegin University, 34794 İstanbul, Türkiye
mms10094@nyu.edu, rana.reda@defi.com.eg, abdelaziz.hussein@ozu.edu.tr

Abstract

Navigating the complexities of Arabic readability prediction requires addressing the language's rich morphology and structural diversity. In the BAREC Shared Task 2025, we participated in all tracks using a stacked ensemble meta learning framework. Our approach combined seven fine-tuned transformer, whose outputs fed into a meta classifier trained on multiple features, including individual predictions, their average, and the average top prediction probabilities. On the blind test set, our ensemble achieved a Quadratic Weighted Kappa (QWK) of 86.4%, demonstrating the effectiveness of integrating diverse transformer encoders for fine grained Arabic readability classification and the potential of meta learning in morphologically rich contexts.

1 Introduction

Arabic readability prediction assesses how difficult a text is for its intended audience, supporting applications such as text simplification (Fang et al., 2025), adaptive learning (Fitrianto et al., 2024), and automated grading (Qwaider et al., 2025). In Arabic, the task is particularly challenging due to the language's morphological richness and wide dialectal variation, and it also plays a crucial role in promoting equitable access to information for readers of varying proficiency levels.

The Balanced Arabic Readability Evaluation Corpus (BAREC) dataset (Elmadani et al., 2025a) heightens this complexity by covering multiple genres from news, literature, educational content, children poems, social media and more other genres that was discussed by them. This diversity introduces significant lexical, syntactic, and stylistic variation, requiring models to capture cues from orthographic patterns to higher level semantics.

In our effort to contribute to this evolving field, we participated in the BAREC Shared Task 2025 (Elmadani et al., 2025b), which focuses on sentence and document level Arabic readability prediction across 19 distinct difficulty levels. The competition comprises three tracks: a Strict track, where only BAREC data is permitted for training; a Constrained track, where the BAREC dataset, SAMER corpus (Alhafni et al., 2024), and SAMER lexicon (Al Khalil et al., 2020) are available; and an Open track, where any external resources may be used.

Our main objective was to assess whether a stacked meta learning system could achieve competitive performance by leveraging the strengths of several fine-tuned transformer models. In this framework, seven transformer based language models served as base predictors, followed by a meta classifier trained on multiple features details of which will be discussed later on to predict the final readability level. We extended the system in the constrained track by incorporating lexical features extracted from the SAMER lexicon which is an arabic readability resource that assigns difficulty levels to individual words, making it possible to estimate text complexity based on its lexical content. In the open track, we also explored a prompt based zero shot approach with GPT 4.1, by feeding the model with a structured annotation guidelines to guide and refine its predictions.

Our stacked meta learning system achieved 2nd place in Track 1 (sentence level) and 2nd in Track 2 (sentence level), but only 7th in document level Track 1 and was not tested in Track 2 due to poor performance. This indicates its strength at the sentence level but limited effectiveness for documents, partly due to a trade off between QWK and accuracy. Employing the LLM for human like annotation was also ineffective.

The paper is organized as follows: §2 reviews related work, §3 presents the dataset, §4 the methodology, §5 the results, §6 the discussion, and §8 the conclusion and future work.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 320–330

2 Related Work

Zalmout et al. (2016) showed that early automatic readability assessment relied on traditional formulas like Flesch Reading Ease (Flesch and Gould, 1949), Flesch Kincaid (Kincaid et al., 1975), and Dale Chall (Dale and Chall, 1948), focusing on surface features such as sentence and word length and vocabulary familiarity. They later extended this by incorporating lexical and syntactic features into SVMs.

For Arabic, Saddiki et al. (2018) conducted the most extensive study, employing a wide range of lexical and syntactic features for L1 and L2 tasks. Their results demonstrated that leveraging L1 features can improve L2 readability prediction, highlighting the benefits of cross task feature sharing.

Ambati et al. (2016) compared syntactic features from incremental CCG and non-incremental phrase parsers, showing that incremental parsing enhanced both accuracy and speed, with further improvements from adding psycholinguistic features. Similarly, Deutsch et al. (2020) found that neural models, ranging from SVMs to BERT, can match or outperform feature augmented systems when trained on sufficient data, suggesting that deep models already capture key readability indicators.

Liberato et al. (2024) introduced a multi model framework for Arabic word level readability (Hazim et al., 2022) and fragment level readability, combining lexicon, frequency, statistical, and transformer based models, and demonstrated that cascaded and aggregation strategies yield stronger results. Recent research further explores deep learning approaches (Lee and Vajjala, 2022; Imperial and Kochmar, 2023) and the use of large language models (LLMs) (Naous et al., 2024; Huang et al., 2024; Marulli et al., 2024), leveraging their advanced language understanding to predict and analyze readability with greater nuance.

Building on prior work, we participated in all three tracks of the shared task, Track 1 (sentence and document level), Track 2 (sentence level), and Track 3 (sentence level) exploring two main directions: **(1)** integrating machine learning models with fine-tuned models to leverage the strengths of both through a stacked meta classifier in Track 1&2, and **(2)** evaluating the capacity of LLMs in Track 3 to emulate human annotation through systematic prompt engineering.

3 Data

3.1 BAREC Dataset

The Balanced Arabic Readability Evaluation Corpus (BAREC) is a large scale dataset for Arabic readability assessment, containing 69,441 manually annotated sentences (over one million words) across 19 readability levels, ranging from kindergarten to postgraduate. It is designed to balance genre, topic, and audience coverage, providing a rich resource for evaluating Arabic text complexity.

The dataset is divided into four subsets: training, validation, and public test splits, which are provided during the development phase, and a private test set, which is used to evaluate the final systems after the development phase concludes.

We conducted thorough evaluations using the validation and public test sets, followed by a final assessment of the system on the blind test set provided for the shared task.

	# Docs	# Sentences	# Words
Train	1,518	54,845	832,743
Dev	194	7,310	101,364
Public Test	210	7,286	105,264
Blind Test	100	3,420	53,052

Table 1: Dataset statistics for the training, validation, public test, and blind test sets.

3.2 SAMER Lexicon

We present the SAMER Lexicon, a 40K lemma leveled readability resource for Arabic. The lexicon comprises 40,000 lemma and part of speech pairs, each annotated with one of five readability levels. This resource offers a standardized reference for assessing lexical difficulty, enabling its integration into a wide range of readability prediction and educational technology applications.

4 Methodology

4.1 Overview

In this paper, we present our submissions for the three tracks of the BAREC shared task. For Tracks 1 and 2, we followed the recommendation from the BAREC main paper, which suggested framing the task as a regression problem to achieve higher QWK scores. Building on this, we fine-tuned multiple transformers and then trained a stacked meta classifier ML model to predict the final readability level based on their outputs. In contrast, Track 3 adopts a fundamentally different approach: we

experimented with LLMs, specifically leveraging the ChatGPT 4.1, to generate predictions directly through prompt based inference.

4.2 Track 1 & Track 2: Stacked Meta Classifier Approach

The preprocessing stage involved removing both kashida and all diacritics from the text. We then fine-tuned several transformer based models, where the input was the complete sentence and the target label was the readability level of that sentence. The outputs from the fine-tuned models were used as inputs to a stacked meta classifier with three feature types: **(1)** raw predictions from each model, **(2)** their average, and **(3)** the average of the top prediction probabilities. We tested these features individually and in combination. For efficiency in deployment, the meta classifier was implemented as a lightweight ML model (classifier or regressor) operating on model predictions rather than raw text.

We experimented with using these features individually and in combination, presenting only the best results achieved through the combination of all three features. To ensure computational efficiency in the final deployment stage, we implemented the meta classifier as a lightweight machine learning model (either a classifier or a regressor) that operates on the predictions of the fine-tuned models.

Added Lexical Features for Track 2 While Track 2 used the same pipeline, we augmented the meta classifier input with three lexical features from the SAMER lexicon. Each word was lemmatized using the CAMeL Tools MSA disambiguator (Obeid et al., 2020), then matched to the SAMER lexicon; if not found, the closest match was selected via edit distance. For each sentence, we calculated **(1)** the most frequent, **(2)** the maximum, and **(3)** the average SAMER level, which were concatenated with the existing meta classifier features to improve prediction accuracy.

We fine-tuned several transformer based models, including AraBERTv02 (Antoun et al., 2020), AraBERTv2 (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021), bert base arabic camelbert msa (Inoue et al., 2021), XLM RoBERTa large (Conneau et al., 2019), bert qarib (Abdelali et al., 2021) and NuSentiment multilingual (Wang et al., 2024), following the regression based setup described earlier.

For the document level setting, we applied the same process at the sentence level, then assigned each document the maximum readability level predicted for any sentence it contained.

4.3 Track 3: LLM Based Approach

In this track, our goal was to emulate human annotation using a powerful LLM by embedding the full Arabic annotation guidelines (Habash et al., 2025) into the prompt. These guidelines define the evaluator's role, describe the 19 readability levels across six linguistic dimensions, and provide examples, constraints, and ACTFL aligned progression from simplest to most complex. By embedding these criteria in the prompt, we guided the LLM to produce annotations consistent with human judgments, enhanced through prompt engineering techniques such as role specification, task definition, criteria conditioning, and strict output formatting.

5 Results

5.1 Sentence Level

5.1.1 Track 1 & 2

We evaluated the performance of the fine-tuned models individually as well as within the meta learning framework. As shown in table 2, **CAMeLBERT-MSA** achieved the highest performance among all base models, with a QWK of 82.8% on the development set and 83.8% on the public test set.

Model	Dev-QWK	Dev-Acc	Test-QWK	Test-Acc
Arabertv2	77.9%	27.0%	78.8%	27.2%
Arabertv02	80.9%	29.9%	81.9%	29.3%
MArabertv2	81.4%	28.1%	82.1%	28.2%
camel_bert_msa	**82.8%**	36.7%	**83.8%**	36.5%
XLM-ROBERTA	80.6%	**38.5%**	81.8%	**39.3%**
bert_qarib	79.9%	26.6%	81.3%	26.0%
Nu_sent	81.1%	27.8%	82.1%	27.9%

Table 2: Performance of base models on the dev and public test sets (Track 1, sentence-level prediction) for QWK and accuracy.

For the ensemble setting, we conducted an extensive series of experiments using a wide range of machine learning classifiers and regressors. As shown in Table 3, the Naïve Bayes models both Gaussian and Categorical consistently yielded the best results. This result was observed when training on the individual predictions of the seven fine-tuned models, and further improved when incorporating the average score across models. We explored all possible combinations of model predictions, and the best performance was achieved when using the predictions from all seven models together. Performance increased even more when we additionally

included the average of the top predicted probabilities for each instance.

Model	Dev-QWK	Dev-Acc	Test-QWK	Test-Acc
Logistic Regression	81.9%	**45.1%**	82.8%	**44.5%**
Linear Regression	82.6%	37.9%	83.8%	39.0%
Random Forest Classifier	81.0%	41.7%	81.5%	41.4%
Random Forest Regressor	81.8%	39.5%	82.7%	39.8%
GaussianNB	**83.9%**	39.2%	**84.9%**	38.1%
CategoricalNB	83.7%	38.1%	**84.9%**	37.6%
Bagging Classifier	80.6%	42.1%	81.1%	41.6%
Bagging Regressor	81.6%	39.9%	82.5%	39.4%

Table 3: Performance of the meta classifier on the dev and public test sets

For track 2, The same ensemble configuration was then applied with the addition of the previously discussed features. As shown in Table 4, this led to only a marginal improvement on the overall performance.

Model	Dev-QWK	Dev-Acc	Test-QWK	Test-Acc
Logistic Regression	81.8%	**45.1%**	82.8%	**44.4%**
Linear Regression	82.7%	38.0%	83.8%	39.1%
Random Forest Classifier	81.7%	44.9%	81.5%	44.7%
Random Forest Regressor	82.0%	40.3%	82.7%	40.6%
GaussianNB	**83.9%**	38.9%	**84.9%**	37.7%
CategoricalNB	83.7%	38.1%	**84.9%**	37.7%
Bagging Classifier	80.9%	44.3%	81.1%	43.6%
Bagging Regressor	81.5%	39.8%	82.5%	39.9%

Table 4: Performance of the meta classifier on the dev and public test sets.

We selected the CategoricalNB model due to its outstanding performance on the public test set and applied it to the blind test. The results are presented in Table 5.

Track	Model	QWK	Acc
Track 1	CategoricalNB	86.4%	39.7%
Track 2	CategoricalNB	86.4%	39.9%

Table 5: Overall performance on Track 1 and Track 2 Blind test.

5.1.2 Track 3

For the LLM trial, even after extensive prompt engineering and providing the ChatGPT 4.1 API with the full BAREC guidelines, performance was poor, achieving only 40.7% QWK on the dev set when predicitng on the sentence level.

5.2 Document Level

For the document level assessment, we applied the previously described approach on Track1; however, it yielded suboptimal results with 69.6% QWK and 34% accuracy. The reasons for this underperformance are examined in detail in the Discussion section. Since the results were unsatisfactory on Track1, we did not extend this approach to Track 2.

6 Discussion

The results show that the stacked meta learner classifier has a strong positive impact compared to individual fine-tuned models, with CategoricalNB achieving slightly better performance than GaussianNB for sentence level predictions. However, this approach did not transfer well to the document level, where higher accuracy is crucial. Regression based models, while yielding high QWK, tend to have lower accuracy, which limits their effectiveness for document level prediction.

Adding the lexical features produced only a marginal improvement of 0.2% in accuracy, indicating limited impact.

For Track 3, using GPT-4.1 with the provided guidelines and a few prompt engineering techniques performed poorly, failing to effectively mimic the human annotation process.

7 Error Analysis

As shown in appendix figures 3 and 4 , the model excels on Classes 10, 8, and 7 but struggles with 6, 1, and 5, often confusing them with neighboring levels. Mid levels are more distinct, while lower levels exhibit significant overlap.

8 Conclusion & Future Work

This shared task provided a valuable opportunity to advance Arabic readability prediction by comparing diverse modeling strategies across three tracks. Our results highlight the effectiveness of a stacked meta learner, which consistently outperformed individual fine-tuned transformer models, with CategoricalNB delivering the best sentence level results. However, the approach proved less effective for document level prediction, where the accuracy QWK trade off in the fine-tuned models.

These findings emphasize the need for models that balance both accuracy and QWK for document level prediction, as well as more impactful feature integration strategies. Future directions include exploring hybrid architectures, leveraging contextual lexical embeddings, and developing advanced prompting or fine-tuning methods for LLMs to better align outputs with human judgments.

Ethics Statement

Although certain authors maintain institutional affiliations with entities linked to the shared-task organizers, the organizers themselves did not contribute to the ideation, construction, or testing of our systems. The entirety of our work was conducted using only those resources that were openly and uniformly distributed to the participant community, with no selective access or special guidance granted.

Acknowledgments

We acknowledge the contribution of the High Performance Computing Center at NYU Abu Dhabi in providing the necessary support for this research.

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.

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.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057, San Diego, California. Association for Computational Linguistics.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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.

Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025. Progressive document-level text simplification via large language models. *arXiv preprint arXiv:2501.03857*.

Ibnu Fitrianto, Cahya Edi Setyawan, and Malikus Saleh. 2024. Utilizing artificial intelligence for personalized arabic language learning plans. *International Journal of Post Axial: Futuristic Teaching and Learning*, pages 30–40.

Rudolf Franz Flesch and Alan J Gould. 1949. The art of readable writing. *(No Title)*.

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.

Chieh-Yang Huang, Jing Wei, and Ting-Hao Kenneth Huang. 2024. Generating educational materials with different levels of readability using llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. 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.

Fiammetta Marulli, Lelio Campanile, Maria Stella de Biase, Stefano Marrone, Laura Verde, and Marianna Bifulco. 2024. Understanding readability of large language models output: an empirical analysis. *Procedia Computer Science*, 246:5273–5282.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing Arabic automated essay scoring with synthetic data and error injection. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 549–563, Vienna, Austria. Association for Computational Linguistics.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Nasser Zalmout, Hind Saddiki, and Nizar Habash. 2016. Analysis of foreign language teaching methods: An automatic readability approach. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 122–130, Osaka, Japan. The COLING 2016 Organizing Committee.

A Full Evaluation Metrices per approach

Tables 6, 7 and 8 present the extended results, including the four metrics reported in the BAREC paper. As noted earlier, both CategoricalNB and GaussianNB consistently achieve high QWK scores relative to the other models. However, on the blind test set, CategoricalNB consistently outperforms GaussianNB.

Model	D-QWK	D-Acc	D-±1	D-Dist	T-QWK	T-Acc	T-±1	T-Dist
Arabertv2	77.9%	27.0%	63.7%	1.41	78.8%	27.2%	64.4%	1.36
Arabertv02	80.9%	29.9%	68.1%	1.31	81.9%	29.3%	69.2%	1.27
MArabertv2	81.4%	28.1%	66.8%	1.36	82.1%	28.2%	67.1%	1.31
camel_bert_msa	**82.8%**	36.7%	**71.5%**	**1.22**	**83.8%**	36.5%	**72.3%**	**1.17**
XLM-RoBERTa	80.6%	**38.5%**	70.6%	1.26	81.8%	**39.3%**	71.5%	1.20
bert_qarib	79.9%	26.6%	66.9%	1.37	81.3%	26.0%	68.2%	1.31
Nu_sent	81.1%	27.8%	66.6%	1.38	82.1%	27.9%	66.7%	1.33

Table 6: Extended performance of base models on the dev (D) and public test (T) sets (Track 1: Sentence Level), including QWK, accuracy, ±1 accuracy, and distribution distance.

Model	D-QWK	D-Acc	D-±1	D-Dist	T-QWK	T-Acc	T-±1	T-Dist
Logistic Regression	81.9%	45.1%	64.7%	0.36	82.8%	44.5%	65.3%	0.34
Linear Regression	82.6%	37.9%	**72.0%**	0.36	83.8%	39.0%	**72.1%**	0.37
Random Forest Classifier	81.0%	41.7%	65.5%	0.31	81.5%	41.4%	66.3%	**0.26**
Random Forest Regressor	81.8%	39.5%	68.9%	0.30	82.7%	39.8%	69.7%	0.29
GaussianNB	**83.9%**	39.2%	66.8%	**0.27**	**84.9%**	38.1%	67.1%	0.30
CategoricalNB	83.7%	38.1%	70.1%	0.34	**84.9%**	37.6%	70.3%	0.34
Bagging Classifier	80.6%	**42.1%**	65.8%	0.29	81.1%	**41.6%**	66.3%	0.30
Bagging Regressor	81.6%	39.9%	68.0%	0.29	82.5%	39.4%	68.7%	0.28

Table 7: Extended performance of ensemble models on the dev (D) and public test (T) sets (Track 1: Sentence Level), including QWK, accuracy, ±1 accuracy, and distribution distance.

Model	D-QWK	D-Acc	D-±1	D-Dist	T-QWK	T-Acc	T-±1	T-Dist
Logistic Regression	81.8%	45.1%	64.7%	0.36	82.8%	44.4%	65.3%	0.33
Linear Regression	82.7%	38.0%	**72.0%**	0.35	83.8%	39.1%	**72.0%**	0.37
Random Forest Classifier	81.7%	44.9%	66.6%	0.30	81.5%	44.7%	67.2%	**0.28**
Random Forest Regressor	82.0%	40.3%	70.1%	0.31	82.7%	40.6%	70.7%	0.30
GaussianNB	**83.9%**	38.9%	66.6%	**0.27**	**84.9%**	37.7%	67.0%	0.30
CategoricalNB	83.7%	38.1%	70.1%	0.34	**84.9%**	37.7%	70.5%	0.33
Bagging Classifier	80.9%	**44.3%**	66.4%	0.31	81.1%	**43.6%**	67.3%	0.32
Bagging Regressor	81.5%	39.8%	69.2%	0.31	82.5%	39.9%	70.1%	0.30

Table 8: Extended performance of ensemble models on the dev (D) and public test (T) sets (Track 2: Sentence Level), including QWK, accuracy, ±1 accuracy, and distributional distance.

B Prompt Details and Example

The following figure 1 illustrates the prompt used in the third track of the shared task, where we explored prompt based zero shot classification using GPT 4.1. In this setting, the model was provided with structured BAREC annotation guidelines to mimic human labeling. Figure 2 presents the guidelines extracted from Habash et al. (2025), which were embedded in the prompt to serve as a rubric or set of criteria for guiding the model in selecting the appropriate readability level.

التعليمات:
أنت خبير لغوي متخصص في تقييم مستوى قابلية القراءة للنصوص المكتوبة باللغة العربية. وظيفتك هي تحليل الجملة المعطاة وتصنيفها وفقًا لمستوى من مستويات القراءة المحددة، والتي تتدرج من (1) إلى (19).

عند تصنيف الجملة، استند بدقة إلى المعايير التالية لكل مستوى، والتي تشمل عدد الكلمات، نوع المفردات، التراكيب النحوية، التصريفات، الدلالات، ومستوى الرمزية أو المجاز.

مستويات القراءة:
{{guidelines}}

الجملة:
{{sentence}}

المطلوب:
- ما هو مستوى قابلية القراءة المناسب لهذه الجملة؟
- أجب باسم المستوى فقط (مثال: "1").

الجواب:
{{generated_response}}

Translation:

Instructions:
You are a linguistic expert specializing in evaluating the readability level of Arabic texts. Your task is to analyze the given sentence and classify it according to one of the specified readability levels, which range from (1) to (19).
When classifying the sentence, carefully rely on the following criteria for each level, which include: word count, vocabulary type, syntactic structures, inflections, semantics, and the degree of symbolism or figurative language.

Readability Levels:
{{guidelines}}

Sentence:
{{sentence}}

Task:
- What is the appropriate readability level for this sentence?
- Answer with the name of the level only (e.g., "1").

Answer:
{{generated_response}}

Figure 1: Prompt example for the Arabic Readability Assessment

المستوى:1، وفق تصنيف ACTFL هو "مبتدئ أدنى". يتكون النص في هذا المستوى من كلمة واحدة فقط. من حيث التهجئة والإملاء، يجب أن تكون الكلمة مكونة من مقطع واحد أو مقطعين. أما من حيث التصريف والاشتقاق، فيُستخدم الفعل المضارع المفرد فقط. في التراكيب النحوية، يُكتفى بكلمة واحدة. أما المفردات، فتشمل اسم جنس، واسم علم بشرط أن يكون متداولًا وبسيطًا من حيث التركيب، وضمير متصل، بالإضافة إلى مفردات تتطابق مع العامية (سامر I)، وتشمل أيضًا الأرقام من 1 إلى 10 سواء كانت بالأرقام العربية أو الهندية. وأخيرًا، من حيث الفكرة والمحتوى، يجب أن تكون الفكرة مباشرة وصريحة وحسية، ولا يحتوي النص على أي نوع من الرمزية.

Translation:

Level: 1, *according to the ACTFL classification, is "Novice Low." At this level, the text consists of only one word. From the perspective of spelling and orthography, the word should contain one or two syllables. In terms of inflection and derivation, only the singular present tense verb is used. Syntactically, a single word suffices. The vocabulary includes common nouns, proper names provided they are simple and widely used, attached pronouns, words overlapping with colloquial usage (SAMER I), as well as the numbers 1 through 10 in either Arabic or Hindi numerals. Finally, in terms of content, the idea must be direct, explicit, and concrete, and the text should not include any form of symbolism.*

Figure 2: Example of the guidelines used in the prompt to differentiate between the 19 different readability levels

C Error Analysis Confusion Matrix

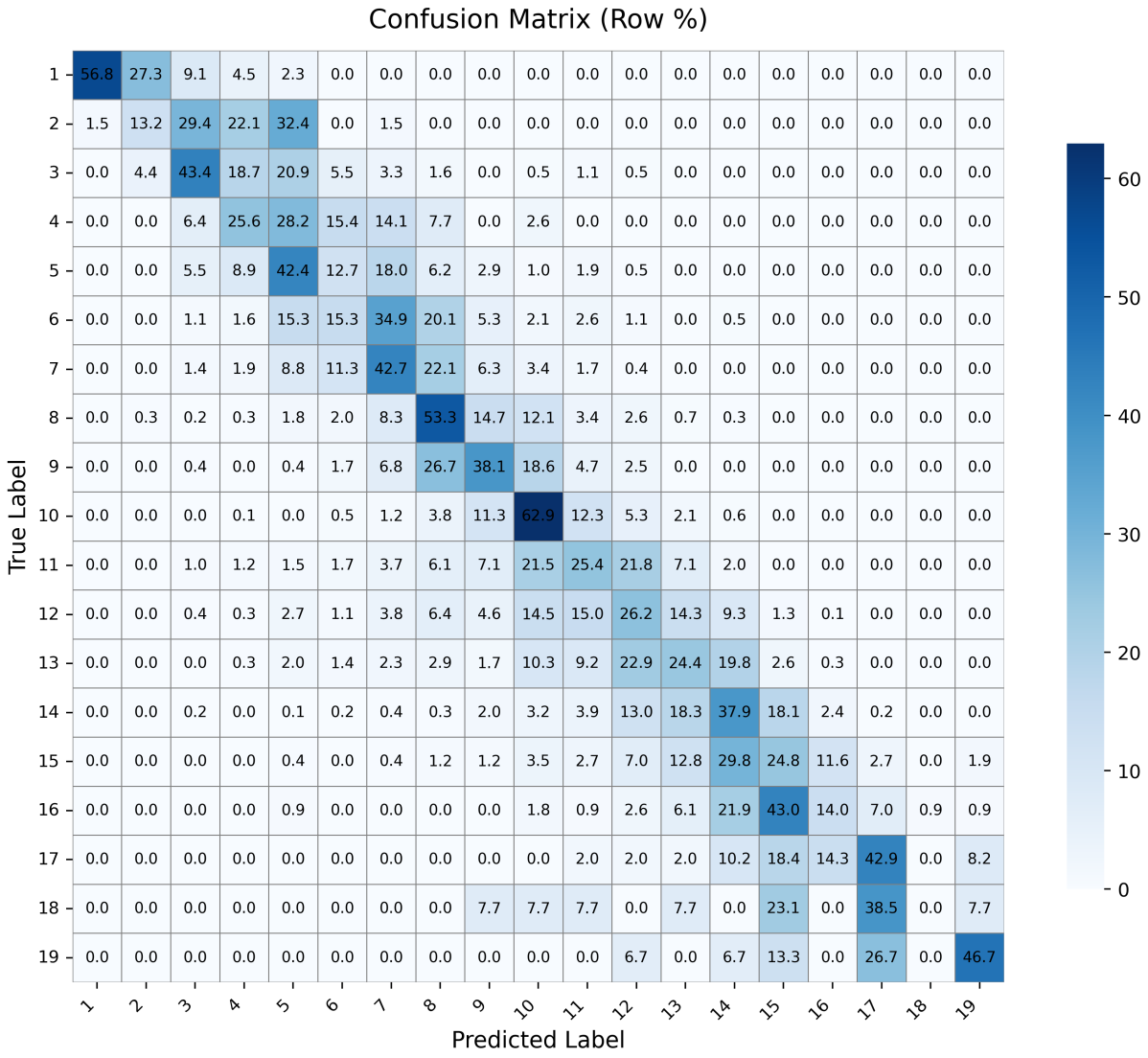

Figure 3: Confusion matrix on the dev set

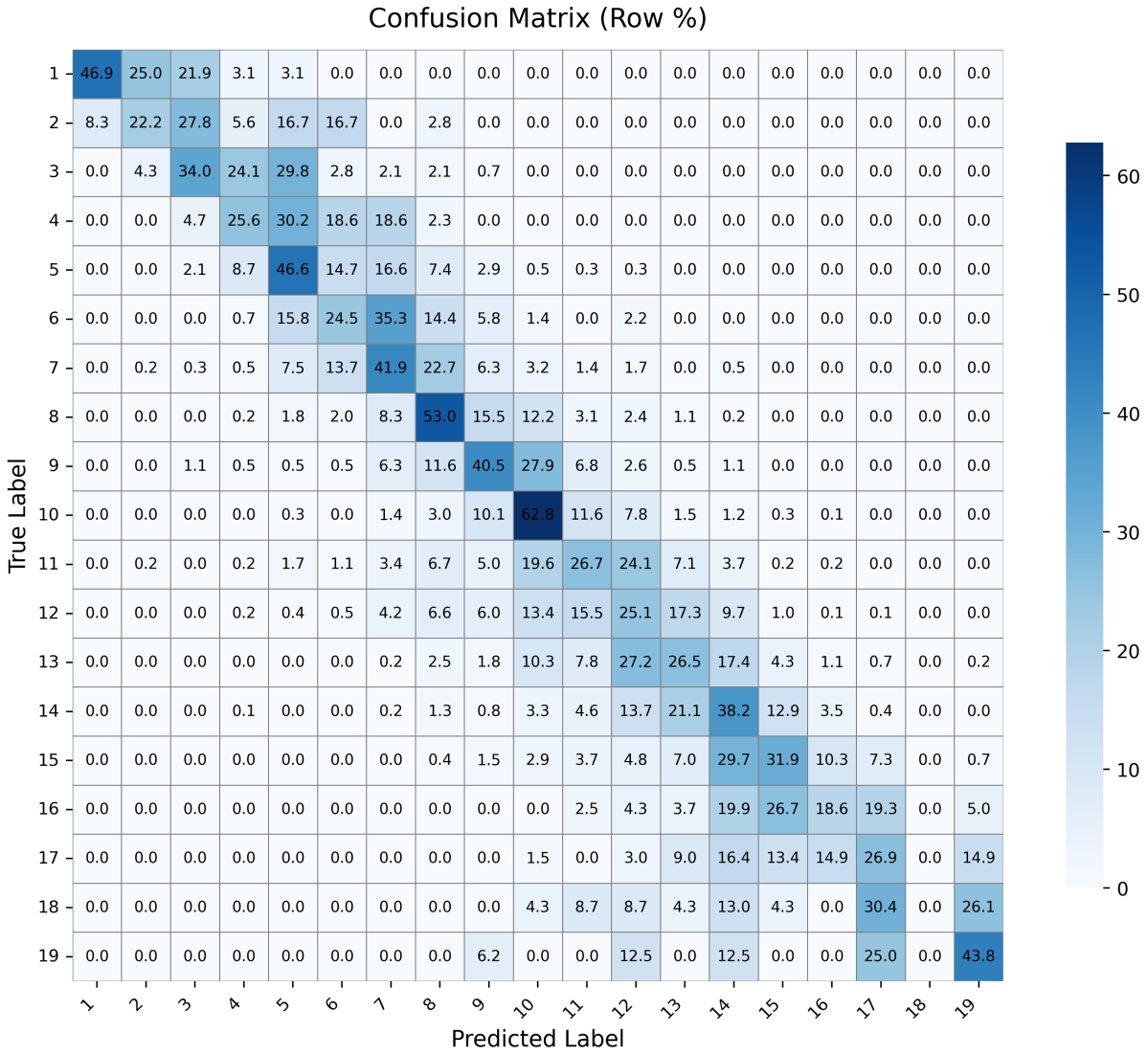

Figure 4: Confusion matrix on the public test set

D Model Hyperparamters

In this paper, we used the same hyperparameters for all models, training on the training set and tuning on the dev set. Each model was trained for 10 epochs with a batch size of 32 for both training and evaluation. We applied a weight decay of 0.01 and used a learning rate of 5e-5. The evaluation strategy was set to run at the end of each epoch, with the best model automatically loaded based on the lowest validation loss, which served as the metric for model selection.