# BAREC Shared Task 2025 on Arabic Readability Assessment

**Khalid N. Elmadani,[1] Bashar Alhafni,[2] Hanada Taha-Thomure,[3] Nizar Habash[1]**
[1]New York University Abu Dhabi
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]Zayed University
{khalid.nabigh,nizar.habash}@nyu.edu
bashar.alhafni@mbzuai.ac.ae, hanada.thomure@zu.ac.ae

## Abstract

We present the results and findings of the BAREC Shared Task 2025 on Arabic Readability Assessment, organized as part of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025). The BAREC 2025 shared task focuses on automatic readability assessment using the BAREC Corpus (Elmadani et al., 2025), addressing fine-grained classification into 19 readability levels. The shared task includes two sub-tasks: sentence-level classification and document-level classification, and three tracks: (1) Strict Track, where only the BAREC Corpus is allowed; (2) Constrained Track, restricted to the BAREC Corpus, SAMER Corpus (Alhafni et al., 2024), and SAMER Lexicon (Al Khalil et al., 2020), and (3) Open Track, allowing any external resources. A total of 22 teams from 12 countries registered for the task. Among these, 17 teams submitted system description papers. The winning team achieved 87.5 QWK on the sentence-level task and 87.4 QWK on the document-level task.[1]

## 1 Introduction

Readability assessment plays a crucial role in education, literacy development, and language learning by ensuring that texts align with a reader's proficiency level. Mismatched readability can lead to less understanding, retention, reading speed, and engagement (DuBay, 2004; Klare, 1963). To address this, text leveling systems have been widely adopted, particularly in early education, to provide structured and measurable progress in reading development (Allington et al., 2015; Barber and Klauda, 2020).

While readability models exist for several languages, many challenges remain, particularly in fine-grained text leveling and resource-scarce languages. Systems like Fountas and Pinnell's 27-level model for English (Fountas and Pinnell, 2006) and Taha-Thomure's (2017) 19-level framework for Arabic demonstrate the importance of detailed readability classification. These fine-grained levels allow for more precise educational applications while being flexible enough to map onto coarser categories for broader applications.

| RL | Grade | Example | |
|---|---|---|---|
| 1 | KG | Majed | ماجد |
| 3 | 1st | The morning of Eid | صباح العيد |
| 6 | 2nd | An idea came to me | جاءتني فكرة |
| 10 | 4th | It was an enjoyable trip! | كانت رحلة ممتعة! |
| 14 | 8th | Definition of Islamic Jurisprudence Principles | تعريف أصول الفقه |
| 17 | Uni | Between lance thrusts and ensign flutters | بين طعن القَنا وخَفْق البُنودِ |

Table 1: Examples by Reading Level (RL) and grade.

The Taha/Arabi21 framework (Taha-Thomure, 2017), which has been used to annotate over 9,000 children's books, plays a central role in our work. Building on this system, the **BAREC** guidelines (Habash et al., 2025) offer standardized, sentence-level readability assessment across a wide range of genres and educational stages – from early childhood to postgraduate levels (see Table 1). For full guidelines in Arabic and English, we refer the reader to Habash et al. (2025).

Arabic, in particular, presents unique challenges for readability assessment due to its rich morphology, extensive lexicon, and highly ambiguous orthography. Unlike English, where well-established readability formulas and datasets exist, Arabic readability research suffers from a lack of standardized resources. This gap limits the development of robust computational models capable of accurately assessing Arabic text difficulty across different proficiency levels.

[1]https://barec.camel-lab.com/sharedtask2025



*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 239–252
November 8-9, 2025 

The BAREC 2025 shared task is organized as part of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025), collocated with EMNLP 2025. A total of 22 teams from 12 countries registered for the shared task. Out of these, 17 teams submitted system description papers which are cited in this paper (see Table 6). This paper provides an overview of the submitted systems and presents their results.

The paper is structured as follows: §2 reviews related work. §3 outlines the sub-tasks and tracks of the shared task. §4 introduces the datasets and evaluation metrics. §5 describes the baselines and provides an overview of the submitted systems. Finally, §6 reports and discusses the results.

## 2 Related Work

**Automatic Readability Assessment** Research on automatic readability assessment has produced a wide range of datasets and resources (Collins-Thompson and Callan, 2004; Pitler and Nenkova, 2008; Feng et al., 2010; Vajjala and Meurers, 2012; Xu et al., 2015; Nadeem and Ostendorf, 2018; Vajjala and Lučić, 2018; Deutsch et al., 2020; Lee et al., 2021). In English, many early datasets were built from textbooks, since their graded structure naturally supports readability evaluation (Vajjala, 2022). Over time, however, copyright limitations and lack of digitized materials pushed researchers to explore alternative sources, such as crowdsourced readability annotations from online platforms (Vajjala and Meurers, 2012; Vajjala and Lučić, 2018), or proficiency exams based on the CEFR framework for L2 learners (Xia et al., 2016).

**Arabic Readability Efforts** Research on Arabic readability has explored text leveling and assessment across several frameworks (Nassiri et al., 2023). Taha-Thomure (2017) proposed a 19-level system for educators, inspired by Fountas and Pinnell (2006), focusing on children's literature. This framework targets full texts, particularly for early education, with 11 of the 19 levels covering up to grade 4, helping teachers match books to students' reading abilities. It defines ten qualitative and quantitative criteria, including text genre, abstractness, vocabulary, dialectal proximity, authenticity, book production quality, content suitability, sentence structure, illustrations, use of diacritics, and word count. The framework was adopted by the Arab Thought Foundation under its Arabi21 initiative, which leveled over 9,000 children's books.

Other approaches applied the CEFR framework (Council of Europe, 2001) to Arabic, including frequency-based word lists from the KELLY project (Kilgariff et al., 2014), manually annotated corpora such as ZAEBUC (Habash and Palfreyman, 2022) and ReadMe++ (Naous et al., 2024), and vocabulary profiling (Soliman and Familiar, 2024). El-Haj et al. (2024) introduced DARES, a dataset derived from Saudi school materials, while the SAMER project (Al Khalil et al., 2020) produced a lexicon with a five-level readability scale, enabling the creation of the first manually annotated Arabic parallel corpus for text simplification (Alhafni et al., 2024). Bashendy et al. (2024) further presented a corpus of Arabic essays annotated for organization and style traits.

Automated Arabic readability assessment has progressed from rule-based models using surface features (Al-Dawsari, 2004; Al-Khalifa and Al-Ajlan, 2010; Hazim et al., 2022) to machine learning approaches incorporating linguistic features (Forsyth, 2014; Saddiki et al., 2018), and script-specific characteristics such as OSMAN (El-Haj and Rayson, 2016). Recent work demonstrates strong performance using pre-trained language models on the SAMER corpus (Liberato et al., 2024).

## 3 Task Description

The BAREC Readability Assessment Shared Task focuses on developing models for fine-grained readability classification using a 19-level framework. Participants built systems to classify texts into these readability levels at both the sentence and document levels.

**Sub-tasks** Participants compete in one or more of the following sub-tasks.

1. **Sentence-Level Classification (Sent)**: Predict the readability level of individual sentences.
2. **Document-Level Classification (Doc)**: Predict the readability level of a document, where a document is a collection of consecutive sentences, and the hardest sentence determines the readability level of the document.

**Tracks** Participants compete in one or more of the following tracks, each imposing different resource constraints:

- **Strict Track (S)**: Models must be trained **exclusively** on the BAREC Corpus (Elmadani

| Split | #Documents | #Sentences | #Words |
|---|---|---|---|
| Train | 1,518 (79%) | 54,845 (79%) | 832,743 (80%) |
| Dev | 194 (10%) | 7,310 (11%) | 101,364 (10%) |
| Test | 210 (11%) | 7,286 (10%) | 105,265 (10%) |
| **All** | 1,922 (100%) | 69,441 (100%) | 1,039,371 (100%) |

Table 2: BAREC Corpus splits.

et al., 2025), ensuring that results are comparable based solely on this dataset.

- **Constrained Track (C)**: Models may use the BAREC Corpus, SAMER Corpus (including document, fragment, and word-level annotations) (Alhafni et al., 2024), and the SAMER Lexicon (Al Khalil et al., 2020).
- **Open Track (O)**: No restrictions on external resources, allowing the use of any publicly available data.

With two sub-tasks and three tracks, the task results in a total of **six possible combinations**. Participants are allowed to compete in multiple sub-tasks and tracks. The goal is to encourage diverse methodological approaches while providing a structured framework for evaluating readability assessment models.

## 4 Shared Task Datasets and Evaluation

In this section, we present the datasets used in different tracks, describe the evaluation metrics, and outline the submission guidelines given to the participants.

### 4.1 Dataset

**BAREC Corpus** The BAREC Corpus (Elmadani et al., 2025) is the main corpus used in the shared task. It consists of 1,922 documents and 69,441 sentences classified into 19 readability levels. The corpus is split into **Train (≃80%)**, **Dev (≃10%)**, and **Test (≃10%)** at the document level. Table 2 shows the corpus splits in the level of documents, sentences, and words. Table 3 shows the label distribution across splits for sentence-level and document-level tasks.

**SAMER Corpus** The SAMER Corpus (Alhafni et al., 2024) consists of 4,289 documents (158K words) and 20,358 fragments classified into three readability levels. We utilize the fragments made available and reported on by Liberato et al. (2024). Table 4 provides an overview of the SAMER corpus statistics, including the Train, Dev, and Test splits.

**SAMER Lexicon** The SAMER Lexicon (Al Khalil et al., 2020) is a 40K-lemma leveled readability lexicon for Modern Standard Arabic (MSA). The lexicon consists of 40K lemma and part-of-speech pairs annotated into five readability levels. The lexicon was manually annotated by three language professionals from different regions in the Arab world. Table 5 shows the readability statistics in the lexicon.

**Blind Test Set** We provide a new blind test set created for this shared task and annotated in the 19 levels of the BAREC framework to evaluate the final results. The blind test set consists of 100 documents and 3,420 sentences.

### 4.2 Evaluation Metrics

Following Elmadani et al. (2025), we treat the Readability Assessment task as an ordinal classification problem and evaluate systems using the following metrics:

- **Accuracy (Acc)** The proportion of cases where predictions exactly match the reference labels in the 19-level scheme (**Acc**$^{19}$). We also report coarse-grained variants: **Acc**$^{7}$, **Acc**$^{5}$, and **Acc**$^{3}$, where the 19 levels are collapsed into 7, 5, and 3 levels, respectively (see Table 3).
- **Adjacent Accuracy (±1 Acc**$^{19}$**)** Also referred to as off-by-1 accuracy, this metric counts predictions as correct if they are either exact matches or differ from the reference by only one level.
- **Average Distance (Dist)** Equivalent to Mean Absolute Error (MAE), it computes the average absolute difference between predicted and reference labels.

| | | | | BAREC Corpus v1 (Sentences) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Level-3** | **Level-5** | **Level-7** | **Level-19** | **All** | | **Train** | | **Dev** | | **Test** | | **Blind Test** | |
| 1 | 1 | 1 | **1-alif** | 409 | 1% | 333 | 1% | 44 | 1% | 32 | 0% | 21 | 1% |
| | | | **2-ba** | 437 | 1% | 333 | 1% | 68 | 1% | 36 | 0% | 21 | 1% |
| | | | **3-jim** | 1,462 | 2% | 1,139 | 2% | 182 | 2% | 141 | 2% | 69 | 2% |
| | | | **4-dal** | 751 | 1% | 587 | 1% | 78 | 1% | 86 | 1% | 28 | 1% |
| | | 2 | **5-ha** | 3,443 | 5% | 2,646 | 5% | 417 | 6% | 380 | 5% | 188 | 5% |
| | | | **6-waw** | 1,534 | 2% | 1,206 | 2% | 189 | 3% | 139 | 2% | 47 | 1% |
| | | | **7-zay** | 5,438 | 8% | 4,152 | 8% | 701 | 10% | 585 | 8% | 296 | 9% |
| | 2 | 3 | **8-Ha** | 5,683 | 8% | 4,529 | 8% | 613 | 8% | 541 | 7% | 263 | 8% |
| | | | **9-ta** | 2,023 | 3% | 1,597 | 3% | 236 | 3% | 190 | 3% | 101 | 3% |
| | | 4 | **10-ya** | 9,763 | 14% | 7,741 | 14% | 1,012 | 14% | 1,010 | 14% | 457 | 13% |
| | | | **11-kaf** | 4,914 | 7% | 4,041 | 7% | 409 | 6% | 464 | 6% | 233 | 7% |
| 2 | 3 | 5 | **12-lam** | 14,471 | 21% | 11,318 | 21% | 1,491 | 20% | 1,662 | 23% | 682 | 20% |
| | | | **13-mim** | 4,039 | 6% | 3,252 | 6% | 349 | 5% | 438 | 6% | 177 | 5% |
| 3 | 4 | 6 | **14-nun** | 10,687 | 15% | 8,573 | 16% | 1,072 | 15% | 1,042 | 14% | 596 | 17% |
| | | | **15-sin** | 2,547 | 4% | 2,016 | 4% | 258 | 4% | 273 | 4% | 171 | 5% |
| | 5 | 7 | **16-ayn** | 1,141 | 2% | 866 | 2% | 114 | 2% | 161 | 2% | 55 | 2% |
| | | | **17-fa** | 480 | 1% | 364 | 1% | 49 | 1% | 67 | 1% | 15 | 0% |
| | | | **18-sad** | 103 | 0% | 67 | 0% | 13 | 0% | 23 | 0% | 0 | 0% |
| | | | **19-qaf** | 116 | 0% | 85 | 0% | 15 | 0% | 16 | 0% | 0 | 0% |
| **Total** | | | | 69,441 | 100% | 54,845 | 100% | 7,310 | 100% | 7,286 | 100% | 3,420 | 100% |

| | | | | BAREC Corpus v1 (Documents) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Level-3** | **Level-5** | **Level-7** | **Level-19** | **All** | | **Train** | | **Dev** | | **Test** | | **Blind Test** | |
| 1 | 1 | 1 | **1-alif** | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **2-ba** | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **3-jim** | 1 | 0% | 1 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **4-dal** | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | 2 | **5-ha** | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **6-waw** | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **7-zay** | 1 | 0% | 1 | 0% | 0 | 0% | 0 | 0% | 1 | 1% |
| | 2 | 3 | **8-Ha** | 1 | 0% | 1 | 0% | 0 | 0% | 0 | 0% | 1 | 1% |
| | | | **9-ta** | 2 | 0% | 1 | 0% | 0 | 0% | 1 | 0% | 0 | 0% |
| | | 4 | **10-ya** | 13 | 1% | 13 | 1% | 0 | 0% | 0 | 0% | 0 | 0% |
| | | | **11-kaf** | 18 | 1% | 10 | 1% | 6 | 3% | 2 | 1% | 1 | 1% |
| 2 | 3 | 5 | **12-lam** | 192 | 10% | 148 | 10% | 25 | 13% | 19 | 9% | 7 | 7% |
| | | | **13-mim** | 204 | 11% | 170 | 11% | 14 | 7% | 20 | 10% | 9 | 9% |
| 3 | 4 | 6 | **14-nun** | 623 | 32% | 489 | 32% | 56 | 29% | 78 | 37% | 24 | 24% |
| | | | **15-sin** | 399 | 21% | 317 | 21% | 46 | 24% | 36 | 17% | 25 | 25% |
| | 5 | 7 | **16-ayn** | 267 | 14% | 207 | 14% | 32 | 16% | 28 | 13% | 20 | 20% |
| | | | **17-fa** | 156 | 8% | 130 | 9% | 9 | 5% | 17 | 8% | 12 | 12% |
| | | | **18-sad** | 12 | 1% | 8 | 1% | 2 | 1% | 2 | 1% | 0 | 0% |
| | | | **19-qaf** | 33 | 2% | 22 | 1% | 4 | 2% | 7 | 3% | 0 | 0% |
| **Total** | | | | 1,922 | 100% | 1,518 | 100% | 194 | 100% | 210 | 100% | 100 | 100% |

Table 3: Sentence-level and document-level splits across BAREC readability levels.

| Split | #Documents | #Fragments | #Words |
|---|---|---|---|
| Train | 2,790 (65%) | 14,256 (70%) | 112,828 (71%) |
| Dev | 607 (14%) | 2,948 (14%) | 22,075 (14%) |
| Test | 892 (21%) | 3,154 (15%) | 23,161 (15%) |
| All | 4,289 (100%) | 20,358 (100%) | 158,064 (100%) |

Table 4: SAMER Corpus splits.

| Level | Type Count |
|---|---|
| Level I | 3,545 (9%) |
| Level II | 3,221 (8%) |
| Level III | 5,510 (14%) |
| Level IV | 10,130 (25%) |
| Level V | 18,281 (45%) |
| **Total** | 40,687 (100%) |

Table 5: SAMER Lexicon distributions

- **Quadratic Weighted Kappa (QWK)** An extension of Cohen's Kappa (Cohen, 1968; Doewes et al., 2023), this measure evaluates agreement between predicted and true labels while penalizing larger misclassifications quadratically, giving higher weight to errors farther from the reference.

We report on **QWK** as the primary metric for ranking systems. We prioritize QWK as it better captures the ordinal nature of readability levels, providing smoother, distance-sensitive penalties for misclassifications compared to the hard thresholds of accuracy-based measures. The other metrics are reported in Appendix A.

### 4.3 Submission Guidelines

The shared task is organized in two phases: development and testing. During the **development phase**, we set up CodaBench (Xu et al., 2022) competitions for all tracks. Participants may either evaluate their systems locally on the BAREC Dev and Test sets,[2] or submit their predictions on the BAREC Test set through the corresponding CodaBench competition for each track. Since the BAREC Test set is publicly available, anyone is welcome to participate in this phase.

In the **testing phase**, we release the Official Blind Test set exclusively to registered participants. Registered teams are required to submit system description papers.

Teams are permitted to participate in all tracks; however, their submissions must adhere to the resource constraints defined for each track. Participation in the constrained track is limited to the use of the SAMER corpus and/or lexicon, while participation in the open track requires the use of external resources.

## 5 Participants and Systems

We received 22 team registrations from 12 countries, of which 17 submitted system description papers. Table 6 lists the participating teams along with their affiliations and the tracks they joined. In total, we received 70 submissions during the development phase and 667 submissions during the testing phase. A detailed breakdown of submissions across tracks is provided in Table 8.

### 5.1 Baselines

We employed three baseline models to compare against the participating systems. All baselines are Arabic-specific BERT-base models fine-tuned on the BAREC Corpus, selected from the suite of models trained by Elmadani et al. (2025). These baselines vary along the following dimensions:

- Pretrained model: AraBERTv02 vs. AraBERTv2 (Antoun et al., 2020)
- Input variant: preprocessing of the BAREC Corpus - **Word** (simple sentence tokenization with diacritics and kashida removal) vs. **D3Tok** (tokenization of words into their base and clitic forms)[3]
- Loss function: Cross-entropy loss (CE) vs. Regression using Mean Squared Error (Reg)

Guided by these design choices, we selected the following three baselines:

[2] https://github.com/CAMeL-Lab/barec-shared-task-2025

[3] Preprocessing with CAMeL Tools (Obeid et al., 2020).

| Team | Affiliation | Sent | | | Doc | | |
|---|---|---|---|---|---|---|---|
| | | S | C | O | S | C | O |
| !MSA (Basem et al., 2025) | MSA University, Egypt | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| AMAR (Saeed et al., 2025) | NYU Abu Dhabi, UAE | ✓ | ✓ | | ✓ | | |
| ANLPers (Sibaee et al., 2025) | Prince Sultan University, KSA | ✓ | | | | | |
| GNNinjas (Elchafei et al., 2025) | Ulm University, Germany | | ✓ | ✓ | | ✓ | |
| LIS (NAIT DJOUDI et al., 2025) | Aix Marseille Université, France | ✓ | | | | | |
| MARSAD (Ibrahim et al., 2025) | Northwestern University, Qatar | ✓ | | | ✓ | | |
| MorphoArabia (Emad Eldin, 2025) | Cairo University, Egypt | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| mucAI (Abdou, 2025) | TUM, Germany | ✓ | | | ✓ | | |
| Noor (Rabih, 2025) | MBZUAI, UAE | ✓ | | | | | |
| PalNLP (Ayesh, 2025) | Cardiff University, UK | ✓ | | | | | |
| Phantoms (Alhassan et al., 2025) | CMU Africa, Rwanda | ✓ | | | | | |
| Pixel (Sapirstein, 2025) | Reichman University, Israel | ✓ | | ✓ | | | |
| Qais (Ahmed, 2025) | IMSIU, KSA | ✓ | | ✓ | | | |
| SATLab (Bestgen, 2025) | UCLouvain, Belgique | ✓ | | | ✓ | | |
| STBW (Trigui, 2025) | (Independent), UAE | ✓ | | | ✓ | | |
| Syntaxa (Bahloul, 2025) | TUM, Germany | ✓ | | | | | |
| ZAI (Nazzal, 2025) | Zayed University, UAE | ✓ | | | | | |

Table 6: List of participating teams, along with their affiliations and the tracks they participated in.

| Team | Score | Features | | | Techniques | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N-gram | Embeds | Morph | ML | PLM | LLM | Ord Loss | Ensemble | Label B. | GNN |
| !MSA | **87.5** | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| AMAR | 86.4 | | | ✓ | | ✓ | ✓ | | ✓ | | |
| mucAI | 85.7 | | | ✓ | | ✓ | | | ✓ | ✓ | |
| STBW | 85.6 | | | ✓ | | ✓ | | ✓ | | | |
| ZAI | 85.5 | | | ✓ | | ✓ | | ✓ | | | |
| Baselines | 84.6 | | | ✓ | | ✓ | | ✓ | | | |
| Syntaxa | 84.3 | | ✓ | ✓ | | ✓ | | | | | ✓ |
| MorphoArabia | 84.2 | | | ✓ | | ✓ | | ✓ | | | |
| MARSAD | 84.1 | | | ✓ | | ✓ | | ✓ | | | |
| Noor | 83.1 | | | ✓ | | ✓ | | | | | |
| Qais | 83.0 | | | ✓ | | ✓ | ✓ | | | ✓ | |
| Phantom | 82.7 | | ✓ | ✓ | | ✓ | | | | | |
| LIS | 82.4 | | | | | ✓ | | | | | |
| SATLab | 82.3 | ✓ | | | ✓ | | | | | | |
| PalNLP | 81.1 | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | |
| GNNinjas | 78.5 | | ✓ | ✓ | | ✓ | | | | | ✓ |
| ANLPers | 73.0 | | ✓ | | | ✓ | | ✓ | | | |
| Pixel | 68.4 | | | ✓ | | ✓ | | | | | |

Table 7: Summary of features and techniques employed by participating teams with their best sentence-level QWK scores. *Embeds* refers to embeddings; *Morph* to morphological segmentation or features; *ML* to non-neural machine learning methods (e.g., SVMs); *PLM* to pre-trained language models; *LLM* to large language models used for prediction or data augmentation; *Ord Loss* to loss functions that account for the ordinal nature of labels (e.g., ordinal log loss, regression); *Label B.* to strategies addressing label imbalance; and *GNN* to graph neural networks.

| Task | Development | | | Testing | | |
|---|---|---|---|---|---|---|
| | S | C | O | S | C | O |
| **Sent** | 63 | 1 | 2 | 221 | 78 | 98 |
| **Doc** | 2 | 1 | 1 | 110 | 83 | 77 |
| | | 70 | | | 667 | |

Table 8: Number of valid submissions during the Development and testing phases across tracks.

- **AraBERTv02+Word+CE** (**Baseline I**): serves as a standard baseline, combining the widely used AraBERTv02 with conventional word-level preprocessing and the standard cross-entropy loss.
- **AraBERTv2+D3Tok+CE** (**Baseline II**): included to assess the effect of linguistically motivated tokenization (D3Tok) on classification performance.
- **AraBERTv2+D3Tok+Reg** (**Baseline III**): motivated by the ordinal nature of readability levels, this baseline explores regression that accounts for the distance between predicted and true labels.

### 5.2 Summary of Submitted Systems

A summary of approaches employed by various teams is provided in Table 7. Most teams built on pre-trained language models (PLMs), often enhanced with morphological features, while a few also incorporated ensembling, label balancing, or ordinal-aware loss functions. Teams such as !MSA, AMAR, and Qais further leveraged large language models (LLMs), and graph neural networks (GNNs) (Zhou et al., 2021) were explored by GNNinjas and Syntaxa, while team Pixel explored vision language models (Dosovitskiy et al., 2021; Rust et al., 2023). Traditional non-neural methods were rare. Overall, the strongest approaches combined PLMs with linguistic features and ensembling. Next we present the system description of the best performing team.

### 5.3 !MSA: Best Performing Team

The pipeline of the winning team, !MSA (Basem et al., 2025), begins by preprocessing the data with D3Tok tokenization (Obeid et al., 2020). They further augment the data differently for each track: upsampling for the strict track, SAMER corpus-based augmentation for the constrained track, and paraphrasing 12k entries from the BAREC corpus using the Gemini API for the open track.[4] Beyond preprocessing, the pipeline is consistent across all tracks. They employ an ensemble of models — AraBERTv2 (Antoun et al., 2020), AraElectra (Antoun et al., 2021), MARBERT (Abdul-Mageed et al., 2021), and CamelBERT (Inoue et al., 2021) — trained with different loss functions including Cross-Entropy, Ordinal Log Loss (Castagnos et al., 2022), Regression (Mean Squared Error), and Conditional Ordinal Regression (Cao et al., 2020). The ensemble combines model outputs via a weighted average, where weights are determined based on each model's confidence scores.

The following section provides a general discussion of the results and analyzes how different approaches impacted performance.

## 6 Results

Tables 9 and 10 show the results in QWK for all tracks in the sentence-level and document-level tasks, respectively. The baselines, highlighted in gray, were trained only on the BAREC corpus, and their scores are reported identically across tracks to facilitate comparison with other teams. Overall, five teams outperformed our strongest baseline. In most cases, however, the additional resources available in the constrained and open tracks did not yield improvements over the strict track. Team **!MSA** achieved the highest QWK scores across all tasks and tracks. In this section, we provide a broad analysis of the overall results. We also report on the other metrics in Appendix A.

### 6.1 General Discussion

Table 7 summarizes the participating teams, their best scores, and the features and techniques they employed. The results show a clear dominance of pre-trained language models (PLMs), which were adopted by nearly all teams. The top performers — !MSA (87.5), AMAR (86.4), and mucAI (85.7) — achieved their results by training on morphologically segmented text and combining PLMs with ordinal-aware loss functions and strategies for addressing label imbalance. Ensembling further boosted performance, particularly for the leading teams, by allowing them to leverage multiple models. Morphological features were widely used across systems, underscoring the importance of morphology-aware approaches in Arabic readabil-

[4] https://ai.google.dev/

| Team | Strict | Constrained | Open |
|---|---|---|---|
| !MSA | **87.5** | **86.6** | **86.4** |
| AMAR | 86.4 | 86.4 | |
| mucAI | 85.7 | | |
| STBW | 85.6 | | |
| ZAI* | 85.5 | | |
| Baseline III | 84.6 | 84.6 | 84.6 |
| Syntaxa | 84.3 | | |
| MorphoArabia | 84.2 | 82.9 | 83.9 |
| MARSAD | 84.1 | | |
| Noor | 83.1 | | |
| Phantom | 82.7 | | |
| Qais | 82.5 | | 83.0 |
| LIS | 82.4 | | |
| SATLab | 82.3 | | |
| Baseline II | 81.5 | 81.5 | 81.5 |
| PalNLP | 81.1 | | |
| Baseline I | 80.5 | 80.5 | 80.5 |
| ANLPers* | 73.0 | | |
| Pixel | 66.2 | | 68.4 |
| GNNinjas | | 78.5 | 77.6 |

Table 9: Performance of participating teams across all tracks in the **sentence-level** task. Scores are reported as QWK (%) and sorted based on the performance on the strict track. * denotes systems that used the dev set for training, making their scores not directly comparable to others.

| Team | Strict | Constrained | Open |
|---|---|---|---|
| !MSA | **87.4** | **84.3** | **82.2** |
| MorphoArabia | 79.9 | 75.5 | 79.2 |
| MARSAD | 79.0 | | |
| SATLab | 77.6 | | |
| mucAI | 73.3 | | |
| Baseline III | 72.6 | 72.6 | 72.6 |
| STBW | 72.5 | | |
| AMAR | 69.6 | | |
| Baseline II | 62.0 | 62.0 | 62.0 |
| Baseline I | 57.7 | 57.7 | 57.7 |
| GNNinjas | | 76.9 | |

Table 10: Performance of participating teams across all tracks in the **document-level** task. Scores are reported as QWK (%) and sorted based on the performance on the strict track.

ity assessment. Team **!MSA** led the leaderboard by integrating all of these components. In contrast, traditional machine learning methods and n-gram features were rarely employed and, when used, did not yield competitive results. More modern approaches such as graph neural networks (GNNs) and vision language models (e.g. team Pixel) also failed to provide significant gains. Large language models (LLMs) were explored by three teams for prediction and data augmentation, but in both cases did not outperform PLM-based systems. Overall, the findings highlight that success in this task depended primarily on combining morphological segmentation with PLMs, ensembling, and ordinal-sensitive modeling.

## 7 Conclusion

In this paper, we presented the framework and results of the BAREC 2025 Shared Task on Fine-Grained Arabic Readability Assessment—the first shared task dedicated to this problem. The task featured two subtasks (sentence-level and document-level) and three tracks (strict, constrained, and open). A new blind test set was created for the evaluation, consisting of 3,420 sentences and 100 documents. In total, 22 teams from 12 countries registered, and 17 submitted system description papers. The strong participation highlights the interest in Arabic readability assessment. Looking ahead, we plan to expand available resources and organize future shared tasks to further advance research in this area.

## Limitations

This work has a few limitations worth noting. First, document-level readability was derived from sentence-level readability under the assumption that the hardest sentence determines the overall document level. While simple, this approach often pushes documents toward higher readability levels, since a single difficult sentence can raise the document's level. Second, we adopted Quadratic Weighted Kappa (QWK) as the primary evaluation metric. However, the choice of the most suitable metric for this task remains an open question.

## Acknowledgments

The **BAREC** project is supported by the Abu Dhabi Arabic Language Centre (ALC) / Department of Culture and Tourism, UAE.

## References


Ahmed Abdou. 2025. mucAI at BAREC Shared Task 2025: Towards Uncertainty Aware Arabic Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

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.

Samar Ahmed. 2025. Qais at BAREC Shared Task 2025: A Fine-Grained Approach for Arabic Readability Classification Using a pre-trained model. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

M Al-Dawsari. 2004. The assessment of readability books content (boys-girls) of the first grade of intermediate school according to readability standards. *Sultan Qaboos University, Muscat*.

Hend S Al-Khalifa and Amani A Al-Ajlan. 2010. Automatic readability measurements of the Arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2 C):103–124.

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.

Ahmed Alhassan, Asim Mohamed, and Moayad Elamin. 2025. Phantoms at BAREC Shared Task 2025: Enhancing Arabic Readability Prediction with Hybrid BERT and Linguistic Features. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Richard L Allington, Kimberly McCuiston, and Monica Billen. 2015. What research says about text complexity and learning to read. *The Reading Teacher*, 68(7):491–501.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mutaz Ayesh. 2025. PalNLP at BAREC Shared Task 2025: Predicting Arabic Readability Using Ordinal Regression and K-Fold Ensemble Learning. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Ahmed Bahloul. 2025. Syntaxa at BAREC Shared Task 2025: BERTnParse - Fusion of BERT and Dependency Graphs for Readability Prediction. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Amelia T. Barber and Susan L. Klauda. 2020. How reading motivation and engagement enable reading achievement: Policy implications. *Policy Insights from the Behavioral and Brain Sciences*, 7(1):27–34.

Mohamed Basem, Mohammed Younes, Seif Ahmed, and Abdelrahman Moustafa. 2025. !MSA at BAREC Shared Task 2025: Ensembling Arabic Transformers for Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. Qaes: First publicly-available trait-specific annotations for automated scoring of arabic essays. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 337–351.

Yves Bestgen. 2025. SATLab at BAREC Shared Task 2025: Optimizing a Language-Independent System for Fine-Grained Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

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.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American*

*Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA. Association for Computational Linguistics.

C. o. E. Council of Europe. 2001. Common european framework of reference for languages: learning, teaching, assessment.

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.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

William H DuBay. 2004. The principles of readability. *Online Submission*.

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).

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. DARES: Dataset for Arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.

Passant Elchafei, Mayar Osama, Mohamed Rageh, and Mervat Abu-Elkheir. 2025. GNNinjas at BAREC Shared Task 2025: Lexicon-Enriched Graph Modeling for Arabic Document Readability Prediction. 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. 2025. 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.

Fatimah Mohamed Emad Eldin. 2025. MorphoArabia at BAREC Shared Task 2025: A Hybrid Architecture with Morphological Analysis for Arabic Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Jonathan Forsyth. 2014. Automatic readability prediction for modern standard Arabic. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.

Irene C Fountas and Gay Su Pinnell. 2006. *Leveled books (k-8): Matching texts to readers for effective teaching*. Heinemann Educational Books.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

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.

Shimaa Ibrahim, Md. Rafiul Biswas, Mabrouka Bessghaier, and Wajdi Zaghouani. 2025. MarsadLab at BAREC Shared Task 2025: Strict-Track Readability Prediction with Specialized AraBERT Models on BAREC. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. 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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Adam Kilgariff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, R. Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

G.R. Klare. 1963. *The Measurement of Readability*. Iowa State University Press.

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.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Anya Amel NAIT DJOUDI, Patrice Bellot, and Adrian-Gabriel Chifu. 2025. LIS at BAREC Shared Task 2025: Multi-Scale Curriculum Learning for Arabic Sentence-Level Readability Assessment Using Pretrained Language Models. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

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.

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).

Ahmad M. Nazzal. 2025. ZAI at BAREC Shared Task 2025: AraBERT CORAL for Fine Grained Arabic Readability. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. 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.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Nour Rabih. 2025. Noor at BAREC Shared Task 2025: A Hybrid Transformer-Feature Architecture for Sentence-level Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. *Preprint*, arXiv:2207.06991.

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.

Mostafa Saeed, Rana Waly, and Abdelaziz Ashraf Hussein. 2025. AMAR at BAREC Shared Task 2025: Arabic Meta-learner for Assessing Readability. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Ben Sapirstein. 2025. Pixels at BAREC Shared Task 2025: Visual Arabic Readability Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Serry Sibaee, Omer Nacar, Yasser Alhabashi, Adel Ammar, Yasser Al-Habashi, and Wadii Boulila. 2025. ANLPers at BAREC Shared Task 2025: Readability of Embeddings Training Neural Readability Classifiers on the BAREC Corpus. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Rasha Soliman and Laila Familiar. 2024. Creating a cefr arabic vocabulary profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies*, 11(1):266–286.

Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling* (معايير هنادا طه لتصنيف مستويات النصوص العربية). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).

Saoussan Trigui. 2025. STBW at BAREC Shared Task 2025: AraBERT-v2 with MSE-SoftQWK Loss for Sentence-Level Arabic Readability. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. 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.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

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.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2021. Graph neural networks: A review of methods and applications. *Preprint*, arXiv:1812.08434.

## A Additional Results

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **87.5** | 43.5 | **76.7** | **1.0** | 64.1 | 69.6 | 76.2 |
| AMAR | 86.4 | 39.7 | 73.2 | 1.1 | 60.8 | 67.8 | 76.1 |
| mucAI* | 85.7 | 50.9 | 75.6 | **1.0** | 65.2 | 69.8 | 76.1 |
| STBW | 85.6 | 33.3 | 73.6 | 1.2 | 57.3 | 66.5 | 74.7 |
| ZAI | 85.5 | 48.8 | 73.3 | **1.0** | 64.4 | 69.3 | 75.8 |
| Syntaxa | 84.3 | 51.0 | 72.0 | **1.0** | 64.4 | 68.7 | 75.4 |
| MorphoArabia | 84.2 | 43.5 | 74.0 | 1.1 | 63.0 | 68.3 | 75.3 |
| MARSAD | 84.1 | 52.0 | 74.0 | **1.0** | 65.9 | 70.6 | 76.1 |
| Noor | 83.1 | 56.1 | 72.5 | **1.0** | 67.0 | 70.5 | 75.8 |
| Phantom | 82.7 | **57.6** | 72.3 | **1.0** | 67.4 | 71.3 | **77.2** |
| Qais | 82.5 | 54.8 | 71.8 | 1.1 | 65.1 | 69.5 | 75.3 |
| LIS | 82.4 | 57.5 | 72.4 | **1.0** | **67.8** | **71.5** | 76.4 |
| SATLab | 82.3 | 25.8 | 63.1 | 1.4 | 47.0 | 59.0 | 69.8 |
| PalNLP* | 81.1 | 33.1 | 69.8 | 1.3 | 57.2 | 63.6 | 72.5 |
| ANLPers | 73.0 | 44.7 | 61.4 | 1.4 | 56.3 | 62.0 | 70.0 |
| Pixel | 66.2 | 38.1 | 53.6 | 1.8 | 48.6 | 54.2 | 65.9 |

Table 11: Performance of participating teams in the **strict** track in the **sentence-level** task. Results are sorted based on the QWK score. * denotes systems that used the dev set for training, making their scores not directly comparable to others.

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **86.6** | 44.9 | **75.4** | **1.0** | **63.0** | **68.7** | 75.6 |
| AMAR | 86.4 | 39.9 | 73.0 | 1.1 | 61.0 | 68.1 | **76.3** |
| MorphoArabia | 82.9 | 30.9 | 70.6 | 1.3 | 53.9 | 62.9 | 72.1 |
| GNNinjas | 78.5 | **50.0** | 67.2 | 1.4 | 61.2 | 66.1 | 74.9 |

Table 12: Performance of participating teams in the **constrained** track in the **sentence-level** task. Results are sorted based on the QWK score.

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **86.4** | 41.3 | **75.1** | **1.0** | 61.7 | 67.3 | 74.5 |
| MorphoArabia | 83.9 | 48.8 | 71.3 | 1.1 | 62.5 | 67.6 | 74.3 |
| Qais | 83.0 | **54.2** | 71.8 | 1.1 | **66.0** | **70.0** | **75.8** |
| GNNinjas | 77.6 | 48.7 | 66.5 | 1.3 | 60.7 | 65.2 | 74.5 |
| Pixel | 68.4 | 41.5 | 56.8 | 1.6 | 50.9 | 56.8 | 65.1 |

Table 13: Performance of participating teams in the **open** track in the **sentence-level** task. Results are sorted based on the QWK score.

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **87.4** | **52** | **94** | **0.6** | **81** | **81** | **93** |
| MorphoArabia | 79.9 | 42 | 90 | 0.7 | 71 | 71 | 92 |
| MARSAD | 79.0 | 36 | 84 | 0.8 | 59 | 60 | 85 |
| SATLab | 77.6 | 39 | 88 | 0.8 | 70 | 71 | 87 |
| mucAI | 73.3 | 36 | 86 | 0.8 | 65 | 66 | 89 |
| STBW | 72.5 | 35 | 85 | 0.8 | 67 | 67 | 90 |
| AMAR | 69.6 | 34 | 79 | 0.9 | 70 | 70 | 89 |

Table 14: Performance of participating teams in the **strict** track in the **document-level** task. Results are sorted based on the QWK score.

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **84.3** | **48** | **91** | **0.6** | **77** | **77** | **94** |
| GNNinjas | 76.9 | 42 | 83 | 0.8 | 60 | 61 | 90 |
| MorphoArabia | 75.5 | 34 | 83 | 0.9 | 64 | 65 | 85 |

Table 15: Performance of participating teams in the **constrained** track in the **document-level** task. Results are sorted based on the QWK score.

| Team | QWK | $Acc^{19}$ | $\pm1$ $Acc^{19}$ | Dist | $Acc^{7}$ | $Acc^{5}$ | $Acc^{3}$ |
|---|---|---|---|---|---|---|---|
| !MSA | **82.2** | **50** | **86** | **0.6** | **70** | **70** | 89 |
| MorphoArabia | 79.2 | 37 | **86** | 0.8 | 65 | 65 | **92** |

Table 16: Performance of participating teams in the **open** track in the **document-level** task. Results are sorted based on the QWK score.