Benchmarking Fine-Grained Readability Explanations for Arabic

Nour Rabih[1] **Khalid N. Elmadani**[1,2] **Hanada Taha-Thomure**[3] **Nizar Habash**[1,2]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[3]Zai Arabic Language Research Centre, Zayed University
nour.rabih@mbzuai.ac.ae, {khalid.nabigh,nizar.habash}@nyu.edu

Abstract

In this work, we introduce BAREC-X, the first explanation-annotated Arabic readability dataset, comprising 1,793 sentences drawn from the Balanced Arabic Readability Evaluation Corpus (BAREC) and labeled by expert annotators with linguistic justifications aligned with the BAREC guidelines. Annotators reached an average inter-annotator agreement of 93.3% under the At-Least-1 metric (73.9% Exact Match at the Category+Reason level, and 81.9% at the Category level). We frame readability explanation as a gold-level-conditioned classification task: given a sentence and its gold readability level, a system must return the linguistic reasons that justify the label. We present the first systematic benchmark for this task on Arabic. We evaluate two contrasting paradigms: (i) eight LLMs under zero- and few-shot prompting, and (ii) a fully interpretable, training-free rule-based system grounded in 67 linguistic features spanning morphology, syntax, vocabulary, and syllabic structure. Gemini 3 and GPT-5.5 achieve the highest exact-match agreement on fine-grained reason prediction (55.8% and 56.4% Category+Reason), while the rule-based system achieves the strongest exact-match performance at the Category level (65.9%). Together, BAREC-X and our benchmark establish a transparent foundation for interpretable Arabic readability explanation.[1]

1 Introduction

Readability assessment plays a key role in education, content creation, and language learning. While extensively studied in English (Fountas and Pinnell, 2006; Heilman et al., 2008; Vajjala and Lučić, 2018), Arabic readability at fine-grained and interpretable levels remains underexplored. Most existing approaches predict readability levels (Al-Khalifa and Al-Ajlan, 2010; Nassiri et al., 2018; Elmadani et al., 2025) but offer little insight into why a text receives a given label, a limitation especially pronounced for Arabic given its rich morphology, syntactic variation, and the scarcity of explanation-oriented resources.

Sentence	عالم غريب المعالم ʕAlm gryb Almʕalm world$_{ms}$ strange$_{ms}$ features$_{mp}$ 'A world of strange features'
Readability	Level 11-kaf (Grade 4)
Linguistic Reason	إضافة لفظية False Idafa

Table 1: Readability explanation example. Transliteration is in the HSB Romanization scheme (Habash et al., 2007).

Large language models (LLMs) have shown promise as direct readability assessors (Trott and Rivière, 2024; Paraschiv et al., 2025), raising the question of whether they can also articulate the linguistic explanations for their judgments. Rule-based systems built on handcrafted linguistic features offer the opposite trade-off: lower flexibility but full transparency and traceability (Lee and Lee, 2020; Pan et al., 2025; Lee et al., 2021; Imperial and Kochmar, 2023). However, no benchmark has compared the two in fine-grained readability explanation for Arabic, since there was no dataset of expert linguistic justifications to evaluate against.

In this work, we frame readability explanation as a gold-level-conditioned classification task: given an Arabic sentence and its gold readability level, a system must select the set of linguistic reasons - drawn from the Balanced Arabic Readability Evaluation Corpus (BAREC) guidelines (Habash et al., 2025) - that justify the label. This formulation isolates the explanation from level prediction and enables direct comparison against expert justifications. An example of the target output is shown in Table 1.

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

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3445–3472
November 6-10, 2026 ©2026 Association for Computational Linguistics

Our contributions are as follows:

- We release **BAREC-X**, the first explanation-annotated Arabic readability dataset: 1,793 sentences drawn from BAREC (Elmadani et al., 2025) and labeled by expert annotators with linguistic justifications aligned with the BAREC annotation guidelines (Habash et al., 2025).[1]
- We present the first systematic benchmark of fine-grained Arabic readability explanation classification, evaluating eight LLMs against a fully interpretable, rule-based system grounded in the BAREC guidelines.

The paper is organized as follows: related work (§2); linguistic background and BAREC guidelines (§3); the BAREC-X dataset, the rule-based system, and LLM prompting approach (§4); experimental setup (§5); results and discussion (§6–§8); and conclusions and future work (§9).

2 Related Work

Automatic readability assessment has been widely studied in NLP, resulting in a variety of datasets, models, and evaluation frameworks across languages. Recent advances in readability assessment have been driven by deep learning. Neural architectures, including recurrent models and hierarchical attention networks, have shown strong performance by learning representations directly from text (Sun et al., 2020). Transformer-based pretrained language models, such as BERT (Devlin et al., 2019), further improved accuracy by capturing contextualized linguistic information (Deutsch et al., 2020; Liu et al., 2025). Despite their effectiveness, these models are typically black-box systems that provide limited insight into the linguistic factors underlying their predictions.

As readability models became more complex, a critical question emerged: how to explain why a text is assigned a particular readability level. This has motivated research on computational methods to improve the interpretability of readability classifiers. One line of work applies post-hoc explanation methods that treat the underlying model as a black box. Model-agnostic approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have been used to explain individual predictions by estimating feature importance scores. While these methods can provide local explanations for complex models, their outputs are often difficult for non-expert users to interpret and may not align cleanly with linguistically meaningful categories (Gilpin et al., 2018). Moreover, such explanations are indirect and depend on the behavior of the underlying model rather than reflecting explicit linguistic justification.

An alternative direction focuses on incorporating linguistically motivated features directly into readability models. Traditional readability formulas are inherently interpretable but limited in expressiveness, while hybrid approaches combine neural representations with handcrafted linguistic features to balance performance and transparency (Deutsch et al., 2020). Although these methods demonstrate that linguistic features can be beneficial, particularly in low-resource settings, their explanatory capacity is often implicit and rarely evaluated against human-provided explanations.

Arabic readability research has additional challenges due to the language's rich morphology, flexible word order, and orthographic variation. Machine learning approaches extended Arabic readability modeling by incorporating morphological, lexical, and syntactic features. For example, Al-Khalifa and Al-Ajlan (2010) demonstrated the effectiveness of SVM-based classifiers using handcrafted features, while more recent work showed that pretrained models such as AraBERT (Antoun et al., 2020) achieve higher predictive accuracy (Berrichi et al., 2024; Liberato et al., 2024). Nevertheless, these neural approaches largely prioritize performance and offer limited transparency into the linguistic drivers of readability decisions.

The development of annotated datasets has played a crucial role in advancing Arabic readability research. Existing resources include CEFR-based datasets for Arabic as a foreign language (Khallaf and Sharoff, 2021; Naous et al., 2023) and school-oriented corpora such as DARES (El-Haj et al., 2024). BAREC (Elmadani et al., 2025) represents a fine-grained Arabic readability dataset, comprising 19 levels spanning early childhood to post-graduate comprehension, with high inter-annotator agreement. BAREC provides detailed linguistic guidelines for annotation (Habash et al., 2025), but its labels do not explicitly encode the justifications behind each decision.

In contrast to previous work that emphasizes prediction accuracy or relies on post-hoc explanations, this study centers readability explanations as a primary objective. We build on BAREC by introducing an explanation-annotated subset in which expert annotators provide explicit linguistic justi-

fications aligned with the BAREC guidelines. We further propose a fully interpretable, rule-based explanation model that mirrors the human annotation process and produces structured, linguistically grounded explanations. By directly evaluating agreement between system-generated and human-provided explanations, our work establishes a transparent baseline for interpretable Arabic readability explanation.

3 Linguistic Background

This section outlines the linguistic foundations of Arabic readability assessment as defined in BAREC guidelines (Habash et al., 2025). BAREC adopts a fine-grained, pedagogically motivated readability framework inspired by Taha-Thomure (2017), in which sentences are assigned to one of 19 ordered readability levels corresponding to progressive educational stages, from early literacy to postgraduate comprehension (Habash et al., 2025). These levels follow the Abjad order of Arabic letters (1-alif, 2-ba, 3-jim, through to 19-qaf) and are designed to provide fine distinctions at lower levels, where readability variation is greatest.

Annotators evaluate sentences across six core linguistic dimensions: spelling and phonology, word count, morphology, syntax, vocabulary, and content. These dimensions jointly define the progression of difficulty across the 19 levels, with different dimensions becoming dominant at different stages of readability development. The same dimensions guide both the annotation of our newly introduced dataset and the design of the experiments proposed in this work.

- **Spelling and Phonology:** Word length and syllabic structure influence difficulty, with more complex syllable patterns associated with higher levels. Diacritics are not considered.
- **Word Count:** The number of unique words serves as an upper bound on readability at early levels.
- **Morphology:** Morphological complexity reflects inflectional and derivational patterns, including clitics, verb forms, plurality, and less frequent constructions.
- **Syntax:** Sentence structure and syntactic relations contribute to difficulty, with non-canonical and complex constructions appearing at higher levels.
- **Vocabulary:** Lexical difficulty is driven by word familiarity and usage, with higher levels introducing rarer, technical, or classical vocabulary.
- **Content:** Conceptual difficulty reflects the level of abstraction and required prior knowledge, ranging from concrete everyday concepts to abstract or domain-specific content.

The complete BAREC annotation guidelines are provided in Appendix A.

4 Methodology

Our methodology has three components: an explanation-annotated dataset, a rule-based explanation model, and LLM-based explanation classification. Building on BAREC, we introduce BAREC-X, a newly annotated subset in which each sentence is labeled with the linguistic reasons that justify its readability level. This dataset enables the direct study and evaluation of explanations in readability modeling.

4.1 Explanation-Annotated Readability Dataset

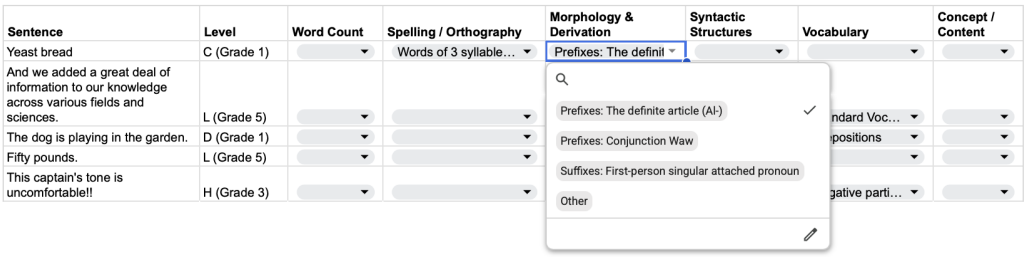
While BAREC provides a strong foundation for sentence-level readability classification, its annotations are limited to readability labels and do not explicitly capture the linguistic justifications behind each decision. To support interpretable readability modeling, we introduce BAREC-X, a newly annotated subset of 1,793 sentences derived from BAREC. These sentences were selected from the inter-annotator agreement (IAA) portion of the corpus to ensure high-quality readability labels.

Annotation Team Three annotators, native Arabic-speaking educators with prior experience in Arabic readability annotation from a related project, were each assigned approximately 800 sentences in total, with around 300 sentences shared across all annotators to enable the measurement of inter-annotator agreement on the selected explanations. Annotator profiles, covering demographic, educational, linguistic, and teaching backgrounds, are listed in Appendix B.

Annotation Batches To facilitate the annotation process, the data was organized into 18 batches of approximately 100 sentences each, which ensured balanced coverage of readability levels while keeping annotation units manageable (see Appendix C, Table 10, for the distribution of levels



(a) Arabic annotation interface.



(b) English translation of the annotation interface.

Figure 1: Screenshot of the Google Sheet annotation interface (a) and its English translation (b).

A1	A2	A3
1	7	13
2	2	2
3	8	14
9	9	9
4	10	15
16	16	16
5	11	17
6	12	18

Table 2: Assignment of annotation batches to annotators A1, A2, and A3.

across batches). Batches were assigned following a partially overlapping design (Table 2): each annotator received five unique batches, and three batches were shared among all three annotators to enable the measurement of inter-annotator agreement.

For each sentence, the annotators were asked to provide explicit linguistic justifications for its assigned readability level from the BAREC corpus. To constrain the annotation space and ensure consistency with the framework, annotators selected from the fixed set of linguistic phenomena that the BAREC guidelines associate with that specific level - rather than from the full inventory of features across all 19 levels. Each sentence may be associated with multiple reasons, reflecting the presence of different factors contributing to readability at that level. Figure 1a illustrates the Google Sheets interface used for annotation, with the English translations shown in Figure 1b. The first two columns contain the sentence and its assigned readability level. The remaining columns correspond to the possible linguistic reasons underlying the assigned level, as defined in the BAREC annotation guidelines (see Appendix A.2). Annotators select one or more applicable reasons for each sentence. An additional notes column is provided for flagging ambiguous or problematic cases and for recording supplementary comments. This process results in aligned sentence-level readability labels and human-provided explanations.

Table 3 shows examples of annotated sentences from five readability levels. Codes follow the format in Appendix E.

Gold Labels For the 300 sentences annotated by all three annotators, the gold set consists of the reasons selected by at least two annotators (majority vote). For the remaining 1,493 sentences, each annotated by a single annotator, the gold set is that annotator's labels. All system evaluations are computed against this gold set, whereas inter-annotator agreement (Section 6) is computed as the average of pairwise agreement between annotators, before unification.

4.2 Rule-Based Explanation Model

We propose a rule-based explanation model designed to mirror the decision process followed by human annotators under the BAREC framework. The model is entirely training-free: it relies exclusively on handcrafted linguistic features and deterministic rules, with no parameters learned from

Sentence	Level	Reasons
الشتاء *'Winter'*	3	M3-1 (proclitic: definite article *Al+*)
ماذا أقرأ؟ *'What should I read?'*	5	S5-3 (basic interrogative particles)
الطقس غائم وبارد وغير مستقر. *'The weather is cloudy, cold, and unstable.'*	8	V8-2 (negation particles)
تُرَى مَنْ يَرْفُضُ أَنْ يَكُونَ حَامِلًا لِأَخْبَارٍ تَجْلِبُ السَعادَةَ؟ *'I wonder, who would refuse to carry news that brings happiness?'*	10	M10-1 (passive voice); S10-2 (*kana* and its sisters)
يَشْرَحُ تَأْثِيرَ المَوارِدِ الطَبِيعِيَة وَالبَشَرِيَة فِي المُجْتَمَعِ. *'It explains the impact of natural and human resources on society.'*	14	V14-2 (general legal, scientific, religious, or political vocabulary)

Table 3: Examples from BAREC-X at five readability levels with their expert-annotated reasons.

data. Given a sentence and a specified readability level, it outputs the linguistic reasons associated with that level by extracting features defined in the BAREC guidelines and returning those whose mapped levels match the input.

Extraction of Linguistic Features Linguistic features corresponding to the dimensions described in Section 3 are extracted using a combination of morphological analysis with CAMeL Tools[2] (Obeid et al., 2020), syntactic parsing with CamelParser[3] (Elshabrawy et al., 2023) , and surface-level measurements. In total, the model operates over 67 linguistically motivated features. These include 25 morphological features, 19 syntactic features, and 19 vocabulary- based features, in addition to surface-level features such as word count and syllable count.

Morphological features capture inflectional and derivational properties such as affixes, clitics, verb tense, voice, and plural forms. Syntactic features encode the presence of specific dependency structures associated with increased sentence complexity. Surface-level features include word count and syllable count, computed in accordance with the BAREC guidelines. Detailed feature definitions and extraction rules are provided in Appendix D.

Explanations Modeling Given a sentence and a target readability level, the model extracts all the defined linguistic features and identifies those that are activated (present in the sentence). Because each feature is deterministically mapped to a fixed readability level in the guidelines, every activated feature contributes one candidate (reason, level) pair. The model then returns the subset of activated reasons whose mapped levels match the input level, yielding the linguistic justification for assigning that label to that sentence. Table 4 illustrates this process: for the sentence عالم غريب المعالم *ςAlm gryb AlmςAlm* 'A world of strange features' with input level 11, only *false idafa* is returned, while features mapped to lower levels (e.g., *broken plural* at level 7, *adjective* at level 2) are filtered out.

Category	Activated Reason	Level
WC	unique word count = 3 (W3-1)	3
SP	syllable count = 3 (O2-1)	2
M	proclitic: definite article *Al+* (M3-1) broken plurals (M7-5)	3 7
V	adjective (V2-3)	2
S	noun + adjective (S2-3) **false idafa** (S11-2)	2 **11**

Table 4: Example of feature competition for عالم غريب المعالم *ςAlm gryb AlmςAlm* 'A world of strange features'. Only activated features are shown. WC=Word Count, SP=Spelling and Phonology, M=Morphology, V=Vocabulary, S=Syntax.

Levels 12-19 are characterized in the BAREC guidelines almost entirely through vocabulary, with progressively rarer, technical, classical, and domain-specialized terms (Appendix A). We capture a subset of these vocabulary phenomena as explicit features (e.g., proper nouns, digits), but the

[2] github.com/CAMeL-Lab/camel_tools, version 1.5.6
[3] github.com/CAMeL-Lab/camel_parser, version 2.0

bulk of the lexical signal at higher levels resists discrete featurization: it lies in the rarity and register of individual lemmas rather than in identifiable categories. For input levels in this range, if no features are activated at the specified level, the model falls back to the vocabulary feature from the guidelines.

4.3 LLM-Based Explanation Classification

We evaluate the ability of LLMs to classify readability explanations directly from the BAREC annotation guidelines. Given a sentence and its target readability level, and the set of candidate reasons associated with that level, the model is prompted to return the set of linguistic reasons that justify the assigned label.

To align the LLM input with the constraints imposed on human annotators (Section 4.1), each prompt includes only the subset of BAREC guidelines associated with the specified readability level. This restricts the model's output space to the same set of candidate phenomena available to annotators, enabling a direct comparison between human and model justifications. The prompt instructs the model to select one or more applicable reasons and to return them in a structured format suitable for evaluation against the BAREC-X annotations.

5 Experimental setup

5.1 Data Splits

To evaluate our systems, we use BAREC-X. All evaluated systems, the rule-based model and the prompted LLMs (zero-, one-, two-, and four-shot), are training-free for this task, so we evaluate the entirety of BAREC-X.

5.2 Evaluation Metrics

We evaluate the explanations from two complementary perspectives: **Category+Reason (CR)** and **Category-only (C)**. The CR setting assesses whether a model correctly identifies the exact linguistic reasons within each category (e.g., specific morphological or syntactic phenomena) provided by human annotators. In contrast, the C setting evaluates performance at a coarser level by considering only the broader linguistic category (e.g., morphology, syntax, vocabulary), regardless of the specific reason selected within that category.

For both evaluation settings, we report three metrics to capture different aspects of alignment with human-provided explanations. First, **Exact Match** measures whether the model's predicted set of reasons exactly matches the annotator-provided set. Second, **Jaccard Similarity** is used to quantify the degree of overlap between predicted and gold sets, providing a graded measure rather than a binary outcome. Finally, **At-Least-1** measures whether the model correctly predicts at least one gold reason, capturing partial but meaningful interpretive overlap. We denote each metric–setting combination with a suffix, e.g., Exact-C and Exact-CR. For the error analysis, we additionally report precision, recall, and F1. Exact Match and At-Least-1 yield a binary score per sentence and are macro-averaged over sentences; Jaccard, precision, recall, and F1 are micro-averaged, i.e., computed over reasons pooled across all sentences.

5.3 Model Architecture and Hyperparameters

Rule-based model does not require traditional training. Each linguistic pattern is explicitly defined by a unique code and linked to a readability level according to the BAREC guidelines; these patterns are treated as discrete codes within the system. During inference, all relevant features are extracted from a sentence, and each activated code contributes to its associated readability level. A complete list of linguistic codes and their corresponding readability levels is provided in Appendix E.

Given a sentence and a target readability level, the model filters the activated codes to return only those whose associated level matches the input, producing the linguistic justification for that label through a fully transparent and deterministic mechanism.

LLMs We evaluate six open-weight LLMs: three Arabic (*ALLaM-7B-Instruct* (Bari et al., 2025), *Fanar-2-27B-Instruct* (TEAM et al., 2026), and *Jais-2-70B-Chat* (Anwar et al., 2025)), and three multilingual models (*Qwen3-30B-A3B-Instruct* (Team, 2025), *Aya-Expanse-32B* (Dang et al., 2024), and *Llama-3.3-70B-Instruct* (Grattafiori et al., 2024)), and two proprietary models: *Gemini-3-Flash* (Team et al., 2025)[4] and *GPT-5.5* (Singh et al., 2026)[5]. Proprietary models are accessed through their official APIs; the total evaluation cost was $51 for GPT-5.5 and $35.2 for Gemini 3, while open-weight models are run locally on 2×A100 80GB GPUs.

[4] *gemini-3-flash-preview* (December 2025)
[5] *gpt-5.5-2026-04-23*

All models are evaluated under four prompting regimes, zero-, one-, two-, and four-shot. To make use of all BAREC-X annotations without leaking gold reasons into the prompt, we adopt a two-fold cross-validation scheme: BAREC-X is split into two folds, and for each test sentence the k in-context exemplars are drawn from the opposite fold. This ensures that every sentence is evaluated under conditions where no part of its annotation has been shown to the model, while still allowing the full dataset to serve as the evaluation set. Shots are fixed rather than resampled per query: for each level and fold, we take the first k annotated sentences of that level from the opposite fold, so all test sentences of a given level and fold, and all models, share the same shots. Each prompt contains: (i) a task description, (ii) the BAREC guideline excerpt for the target level, (iii) k shots (for $k > 0$), and (iv) the target sentence. Models are instructed to return reasons in a structured JSON array format. The full prompt template is provided in Appendix F.

6 Results

6.1 Inter-Annotator Agreement

To assess the quality and consistency of the explanation annotations in the newly introduced interpretable subset, we compute inter-annotator agreement (IAA) on the 300 selected sentences. Since each sentence may be associated with multiple explanation labels, agreement is evaluated using the same metrics described in Section 5.2. Agreement is computed pairwise across annotators.

We report IAA results under both the **CR** and **C** evaluation settings, capturing agreement at fine-grained and coarse-grained levels of explanation. The resulting scores provide an empirical measure of the reliability of the annotation and validate the quality of the explanation labels. Detailed agreement statistics are presented in Table 5.

The inter-annotator agreement results demonstrate strong consistency in the explanation annotations, supporting the reliability of the proposed dataset. Agreement is consistently higher under the coarse-grained C setting than under the finer-grained CR setting, as expected. Substantial Exact Match and Jaccard scores are observed across annotator pairs, while the At-Least-1 metric exceeds 90% in all cases, indicating near-universal agreement on at least one key linguistic factor. These results confirm that the annotation guidelines enable consistent interpretive judgments and that the dataset is suitable for evaluating systems for readability explanation classification.

Annotator Pair	Exact		Jaccard		At-Least-1	
	C	CR	C	CR	C	CR
A1-A2	79.5	68.4	77.9	68.0	94.3	91.3
A2-A3	85.5	82.8	84.4	80.2	97.0	95.3
A1-A3	80.8	70.4	79.5	70.9	95.0	93.3
Average	81.9	73.9	80.6	73.0	95.4	93.3

Table 5: Pairwise inter-annotator agreement (IAA). All values are reported as percentages (%).

6.2 Benchmark Results

Table 6 reports agreement between system-generated and human-provided explanations on BAREC-X under three metrics (Exact, Jaccard and At-Least-1) and two settings (Category-only, C, and Category+Reason, CR; see Section 5.2).

At 4-shot, **Gemini 3** achieves the strongest agreement on both At-Least-1 metrics (C: 89.5%, CR: 87.6%), while **GPT-5.5** narrowly leads on Exact-CR (56.4% vs. Gemini's 55.8%) and the **rule-based** system attains the highest Exact-C score (65.9% vs. Gemini's 64.8%). The three systems are tightly clustered at the top: no single approach dominates across all metrics, and the rule-based system remains competitive with frontier LLMs on category-level agreement despite using no learned parameters. Its main gap is on Exact-CR (40.9%, roughly 15 points below Gemini and GPT), indicating that while the rule-based system reliably identifies the correct linguistic dimension, it is less precise at naming the exact phenomenon within that dimension. Among the open-weight Arabic and multilingual models, **Llama-3.3** and **Fanar-2** are the strongest 4-shot performers, both clearing 60% Exact-C and approaching the frontier models on At-Least-1. **Aya-Expanse** and **Qwen3** follow at mid-range. **Jais-2** and **ALLaM** perform poorly throughout.

Effect of few-shot prompting. Moving from zero- to four-shot prompting improves performance for most models. Averaged across all eight LLMs, gains are modest (**+13.3** Exact-C, **+14.3** Exact-CR, **+3.1** At-Least-1-C, **+6.5** At-Least-1-CR), held down by **ALLaM** and **Jais-2**, which behave differently from the rest and regress with more shots. Excluding these two outliers, the average gain rises substantially to **+20.0** Exact-C, **+19.6** Exact-CR, **+9.4** At-Least-1-C, and **+11.4** At-Least-1-CR,

Model	Shots	Exact		Jaccard		At-Least-1	
		C	CR	C	CR	C	CR
Gemini 3	0	48.9	31.3	54.2	47.4	78.5	77.2
	1	55.0	40.7	60.5	53.3	86.2	84.7
	2	60.4	50.1	63.4	56.8	88.2	86.4
	4	64.8	55.8	**66.7**	**60.3**	**89.5**	**87.6**
GPT-5.5	0	56.3	47.1	57.1	51.7	73.8	72.1
	1	61.9	53.1	62.4	56.4	78.6	76.7
	2	63.2	55.2	63.0	57.2	80.0	78.0
	4	64.1	**56.4**	63.9	58.1	80.3	78.5
ALLaM	0	13.3	5.6	17.5	8.4	24.1	13.9
	1	2.5	1.8	2.5	1.8	3.5	2.8
	2	2.4	2.1	2.7	2.3	3.5	3.2
	4	4.3	3.5	4.8	4.0	5.9	5.2
Aya	0	35.7	24.8	39.8	27.2	69.3	55.0
	1	45.8	34.0	44.6	33.7	69.6	59.6
	2	47.3	35.9	48.8	36.4	75.2	64.1
	4	50.9	41.2	51.7	39.6	78.0	67.3
Fanar-2	0	29.2	21.0	34.9	23.4	67.8	56.8
	1	57.0	44.6	50.7	38.3	74.5	63.9
	2	58.7	47.7	51.4	40.2	74.0	64.7
	4	61.1	49.2	53.5	42.3	75.0	65.9
Jais-2	0	34.8	25.6	33.0	23.0	49.2	39.2
	1	17.2	13.3	18.9	15.1	22.6	19.3
	2	26.8	21.2	27.0	21.0	33.3	28.1
	4	30.0	24.9	28.7	23.2	35.9	31.2
Llama-3.3	0	42.8	33.1	41.1	30.0	64.0	53.3
	1	55.3	40.4	51.5	38.7	71.3	62.4
	2	57.3	41.7	53.9	40.2	75.9	65.8
	4	60.3	44.3	56.4	43.1	77.6	68.8
Qwen3	0	16.9	10.0	33.9	20.5	63.7	47.9
	1	43.7	32.9	42.7	30.9	67.2	56.7
	2	45.4	35.5	46.1	34.7	72.4	62.6
	4	48.9	37.9	47.3	35.7	72.8	63.0
Rule-based	–	**65.9**	40.9	64.0	50.0	78.2	74.7

Table 6: Agreement (%) between system-generated and human-provided explanations on BAREC-X. **C**: Category-only; **CR**: Category+Reason. Best per column in bold.

Category	C			CR		
	Gem	GPT	Rules	Gem	GPT	Rules
Overall	**80.0**	78.0	78.1	**75.2**	73.5	66.6
Spelling & Phon.	70.1	**71.8**	53.1	69.8	**71.4**	51.9
Word Count	54.1	31.3	**71.0**	54.1	31.3	**71.0**
Morphology	**86.4**	84.0	85.0	**84.4**	82.0	81.5
Syntax	**79.0**	76.1	58.4	**69.6**	68.4	49.8
Vocabulary	**88.1**	86.1	87.6	**82.9**	81.4	69.7
Content	16.0	**22.2**	0.0	13.8	**17.5**	0.0

Table 7: Per-category micro F1 (%) on BAREC-X for the two top LLMs (4-shot) and the rule-based system (**Rules**). **C**: Category-only; **CR**: Category+Reason.

Level	*n*	C			CR		
		Gem	GPT	Rules	Gem	GPT	Rules
1	4	**87.5**	85.7	83.3	78.8	**85.7**	83.3
2	12	40.0	54.5	**66.7**	31.3	41.7	**42.1**
3	34	80.5	**82.4**	73.7	79.6	**81.3**	73.3
4	10	**82.4**	**82.4**	80.0	**77.8**	75.7	66.7
5	49	59.3	**73.7**	40.9	59.3	**73.7**	40.9
6	55	63.2	**68.3**	67.3	63.2	**68.3**	67.3
7	134	74.0	**83.1**	68.8	70.9	**78.9**	65.6
8	130	81.7	**82.4**	76.9	80.5	**81.4**	73.9
9	61	90.5	**94.6**	68.0	90.5	**94.6**	68.0
10	209	**86.1**	84.7	72.0	75.9	**76.4**	66.1
11	89	57.4	26.8	**63.5**	55.2	23.3	**60.4**
12	423	77.1	66.3	**78.6**	**72.7**	62.8	69.3
13	93	74.3	**74.9**	60.6	66.4	**69.0**	34.8
14	337	**96.6**	96.0	95.6	87.5	**87.8**	68.4
15	75	73.6	64.8	**86.4**	73.6	64.8	**86.4**
16	47	**97.9**	96.8	**97.9**	**85.0**	84.6	70.3
17	23	91.3	**93.6**	91.7	91.3	**93.6**	91.7
18	3	80.0	80.0	**100.0**	80.0	80.0	**100.0**
19	5	90.9	90.9	90.9	90.9	90.9	90.9

Table 8: Per-level micro F1 (%) on BAREC-X for the two top LLMs (4-shot) and the rule-based system (Rules). *n*: number of sentences per level.

confirming that four-shot prompting reliably helps models that can productively use in-context examples for this task.

7 Error Analysis

Tables 7 and 8 break performance down by linguistic category and by readability level. Four patterns emerge.

Strengths of the rule-based system. The rule-based system leads on **Word Count** (71.0% F1 vs. 54.1% for Gemini and 31.3% for GPT), where deterministic counting beats holistic judgment, and it matches Gemini on **Morphology** (85.0% vs. 86.4% C-F1) and **Vocabulary** (87.6% vs. 88.1%). The per-level results agree. The rule-based system leads at level 2, where counting dominates the guidelines, and at level 11, where explicit rules cover false *idafa* and relative pronouns. At level 11, GPT-5.5 collapses (26.8% C-F1), with low precision and low recall alike (Appendix G, Table 16), indicating that it selects the wrong phenomena rather than abstaining.

Strengths of LLMs. Gemini and GPT lead on **Syntactax** (79.0%/76.1% vs. 58.4% C-F1) and **Spelling & Phonology** (70.1%/71.8% vs. 53.1%). The rule-based system inherits errors from the dependency parser and morphological disambiguator, whereas LLMs recognize constructions holistically. Rule coverage compounds the problem: several phenomena are not implemented (Appendix E), such as conjoined sentences at level 5 and emotion vocabulary and oath constructions at level 9. These

are exactly the levels where the rule-based system is weakest (40.9% and 68.0% vs. 73.7% and 94.6% for GPT).

Right category, wrong reason. The rule-based system shows the largest drop from C to CR (78.1% to 66.6% overall), concentrated in Syntax, Vocabulary, and levels 13–16. At level 14, for example, it nearly matches Gemini under C (95.6% vs. 96.6%) but falls to 68.4% under CR (vs. 87.5%). This is a consequence of the vocabulary fallback (Section 4.2): when no feature fires at levels 12–19, the system defaults to a vocabulary reason, which lands in the correct category but often names the wrong phenomenon within it. The same fallback helps at level 15, where it likely coincides with the dominant gold reason (86.4%, best overall). The LLMs show a smaller version of this drop at the higher levels, suggesting that fine-grained vocabulary reasons (e.g., SAMER bands) are hard to distinguish without access to the underlying lexicon.

Content remains unsolved. **Content** is the hardest category for every system: Gemini reaches 16.0% C-F1 and GPT 22.2%, while the rule-based system scores 0% by construction. Content reasons concern symbolism, abstraction, and required prior knowledge (Appendix D, Table 15), which resist featurization and which LLMs identify unreliably without grounding.

8 Discussion

This work demonstrates that Arabic readability explanation can be modeled in a linguistically faithful manner when annotation, system design, and evaluation are tightly aligned with the BAREC framework. For dimensions that the BAREC guidelines define through explicit linguistic phenomena- Word Count, Morphology, and Vocabulary - deterministic rules are competitive with or superior to large proprietary models, demonstrating that interpretability need not come at the cost of accuracy when the annotation framework is itself linguistically explicit.

The two paradigms exhibit complementary error profiles. The rule-based system trails LLMs on Syntax and Spelling & Phonology, where it inherits errors from upstream morphological and syntactic analyzers, while LLMs lag on Word Count, where deterministic counting trivially beats holistic judgment. The clearest open challenge is Content - all systems struggle here - reflecting the difficulty of identifying abstraction, cultural knowledge, and conceptual density without grounding context. Together, these patterns suggest that the most promising direction is not to choose between paradigms but to route each linguistic dimension to the system best suited to handle it.

Beyond model comparison, the task itself is useful. A system that can articulate *why* a sentence receives a given readability level has direct applications in education and curriculum design: teachers can identify the specific linguistic obstacles a text presents to a target grade, content creators can target rewrites at the phenomena driving difficulty, and Arabic language learners can receive feedback grounded in concrete grammatical and lexical features rather than opaque numeric scores. The high At-Least-1 agreement observed across systems (89.5% for Gemini 3, 78.2% for the rule-based system) indicates that even when systems do not exactly match the human-provided set of reasons, they frequently capture a core justification aligned with expert judgments - sufficient for many educational and diagnostic use cases where partial but faithful explanations are valuable.

9 Conclusions and Future Work

We introduced the first benchmark for fine-grained Arabic readability explanation, framing the task as a gold-level-conditioned classification task. We also released BAREC-X, the first explanation-annotated Arabic readability dataset (1,793 sentences with expert explanations and 93.3% At-Least-1 agreement), and evaluated two paradigms: instruction-tuned LLMs and a fully interpretable, training-free rule-based system built on 67 linguistic features aligned with the BAREC guidelines.

Results show complementary strengths rather than a single winner. Gemini 3 and GPT-5.5 achieve the best fine-grained exact-match scores (55.8% and 56.4% Exact-CR), while the rule-based system attains the highest category-level score (65.9% Exact-C) without learned parameters. Rule-based methods perform best when guidelines are linguistically explicit, whereas LLMs excel in holistic recognition settings.

Future work includes expanding rule-based coverage and analyzer robustness, developing hybrid systems that combine rules and LLMs, and scaling BAREC-X to support training-based approaches and broader evaluation of higher readability levels and content-driven phenomena.

Acknowledgments

The BAREC project is supported by the Abu Dhabi Arabic Language Centre (ALC) / Department of Culture and Tourism, UAE. This research is in part with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program. We also acknowledge the support of the High Performance Computing Center at New York University Abu Dhabi. Finally, we are deeply grateful to our outstanding annotation team: Sawsan Tannir, Samar Zeino, and Zeina Zeino.

Limitations

Despite its strengths, this work has several limitations. Our rule-based system implements 67 linguistic features aligned with the BAREC guidelines, but does not yet cover the full inventory of phenomena described in the framework. Certain fine-grained constructions and rare patterns are not explicitly modeled, particularly at higher readability levels where the guidelines characterize complexity through abstract vocabulary and content properties that resist discrete featurization. Expanding coverage to additional phenomena would likely improve agreement.

The accuracy of the extracted features also depends on the reliability of the underlying morphological disambiguator and dependency parser. Errors in these tools propagate to the feature representations and consequently to the generated explanations. While we rely on well-established resources, further validation of the feature extraction pipeline is necessary to fully isolate the contribution of the rule-based approach from upstream noise.

Our LLM evaluation is a pilot study. We benchmark eight models under a single shared prompt structure across zero-, one-, two-, and four-shot regimes, and we acknowledge that more extensive prompt engineering, particularly tuning the language of the instructions, and the structure of in-context exemplars, could substantially improve performance, especially for the open-weight Arabic models. We restrict ourselves to a uniform setup to enable fair cross-model comparison, and the reported numbers should be read as a lower bound on what well-tuned prompts could achieve. The set of models evaluated, while broad, is also not exhaustive; future work should extend the benchmark to additional Arabic-specialized and reasoning-tuned LLMs as they become available.

Ethical Considerations

The annotation process is conducted with transparency and fairness, with multiple annotators involved to mitigate biases and ensure reliability. All annotators are paid fair wages for their contribution.

The annotated dataset is openly accessible to promote transparency, reproducibility, and collaboration in Arabic language research.

The accuracy of the systems evaluated in this work remains limited, particularly in fine-grained reason prediction and in content-level phenomena. In educational or assessment settings, incorrect or incomplete explanations could mislead teachers, learners, or curriculum designers about the actual linguistic obstacles a text presents. We therefore recommend cautious use of these systems as supportive tools alongside expert human judgment, rather than as standalone authorities on text complexity.

We used AI writing assistance within the scope of "Assistance purely with the language of the paper" described in the ACL Policy on Publication Ethics.

References

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.

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.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan

AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Safae Berrichi, Naoual Nassiri, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2024. Exploring the impact of deep learning techniques on evaluating arabic l1 readability. In *Artificial Intelligence, Data Science and Applications*, pages 1–7, Cham. Springer Nature Switzerland.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

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

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.

Khalid Elmadani, Nizar Habash, and Hanada Taha. 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.

Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. CamelParser2.0: A State-of-the-Art Dependency Parser for Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

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

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

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.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. *arXiv preprint arXiv:2103.04386*.

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.

Bruce W. Lee and Jason Lee. 2020. LXPER index 2.0: Improving text readability assessment model for L2 English students in Korea. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–24, Suzhou, China. Association for Computational Linguistics.

Juan Piñeros Liberato, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2024. Strategies

for arabic readability modeling. *arXiv preprint arXiv:2407.03032*.

Fengkai Liu, Tan Jin, and John SY Lee. 2025. Automatic readability assessment for sentences: neural, hybrid and large language models. *Language Resources and Evaluation*, pages 1–32.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2023. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. *arXiv preprint arXiv:2305.14463*.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018. Modern standard arabic readability prediction. In *Arabic Language Processing: From Theory to Practice*, pages 120–133, Cham. Springer International Publishing.

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.

Wenjing Pan, Xia Li, Xiaoyin Chen, and Rui Xu. 2025. Textual form features for text readability assessment. *Natural Language Processing*, 31(3):800–841.

Andrei Paraschiv, Mihai Dascalu, and Marina Solnyshkina. 2025. Assessing the readability of russian textbooks using large language models. *Information*, 16(12):1071.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Yuxuan Sun, Keying Chen, Lin Sun, and Chenlu Hu. 2020. Attention-based deep learning model for text readability evaluation. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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

FANAR TEAM, Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Asim Ersoy, Masoomali Fatehkia, Mohammed Qusay Hashim, and 18 others. 2026. Fanar 2.0: Arabic generative ai stack. *Preprint*, arXiv:2603.16397.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

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.

A BAREC Annotation Guidelines

A.1 BAREC Guidelines

مستوى بارق	صف	ACTFL	عدد كلمات	تهجئة وإملاء	تصريف واشتقاق	تراكيب نحوية	مفردات	فكرة ومحتوى
أ	روضة-1	مبتدئ أدنى	1	• كلمات من مقطع واحد أو مقطعين	• الفعل المضارع المفرد	• كلمة واحدة	• اسم جنس • اسم علم (متداول بسيط تركيبيا) • ضمير منفصل • مفردات متطابقة مع العامية - سامر I • الأرقام (العربية أو الهندية) 1-10	• فكرة مباشرة وصريحة وحسية. • لا رمزية في النص.
ب	1	مبتدئ أدنى	≤2	• كلمات من 3 مقاطع		• جملة اسمية (هو يلعب) • إضافة حقيقية (باب البيت) • صفة وموصوف (باب كبير)	• فعل • صفة • مفردات متشابهة مع العامية - سامر I • العدد الأصلي بالأحرف • الأسماء الخمسة: أبو، أخو	
ج		مبتدئ متوسط	≤4	• كلمات من 3 مقاطع	• سوابق: ال التعريف • سوابق: واو العطف • لواحق: ضمير المتكلم المفرد المتصل	• بدل كل: (صديقي أحمد) • بدل إشارة: (هذا البيت)	• مفردات فصيحة شائعة - سامر I • اسم الإشارة المفرد • الأرقام (العربية أو الهندية) 11-100	
د		مبتدئ متوسط	≤6	• كلمات تستخدم مد الألف (أ)	• الفعل المضارع الجمع • سوابق: حروف جر متصلة • ظرف منون	• جملة فعلية بدون مفعول به • جار ومجرور	• حروف الجر	
ه	2	مبتدئ أعلى	≤8	• كلمات من 4 مقاطع	• لواحق: ضمير متصل مفرد أو جمع • المثنى (في الأسماء والصفات) • جمع المؤنث السالم	• جملة فعلية مع مفعول به واحد اسم • جمل معطوفة • أدوات استفهام أساسية: ماذا، متى، من، أين، ما، كيف • صيغة التعجب "ما أفعل"	• العدد الترتيبي • الأرقام (العربية أو الهندية) 101-1,000 • اسم اشارة مثنى، جمع	• المحتوى من حياة القارئ. • لا رمزية في النص.
و		مبتدئ أعلى	≤9	• كلمات من 5 مقاطع	• الفعل الماضي المفرد والجمع • جمع مذكر سالم	• جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	• مفردات فصيحة - سامر I	
ز		متوسط أدنى	≤10	• كلمات من +6 مقاطع • أفعال/أسماء معتلة الآخر	• الفعل الماضي المثنى •الفعل المضارع المثنى • فعل الأمر المفرد • لواحق: ضمير المثنى المتصل • جمع التكسير • واو القسم (والله)	• مفعول فيه (ظروف زمان ومكان) • حال • أداة الاستفهام هل	• مفردات فصيحة شائعة - سامر II	• بعض الرمزية أو عدم التصريح المباشر بكل المقصود في الجملة
ح	3	متوسط أدنى	≤11		• فعل الأمر الجمع • نون النسوة في الأسماء والأفعال • سوابق أخرى: سين الاستقبال، واو الاستئناف، فاء العطف • أدوات ربط (ثم، حتى، أو، أم، لكن، أمّا)	• المفعول المطلق • المفعول لأجله • المفعول معه • جملة فعلية تتعدى إلى مفعولين	• مفردات فصيحة - سامر I و سامر II • أحرف النفي • الأرقام (العربية أو الهندية) 1,001-1,000,000	• بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة
ط		متوسط أوسط	≤12		• فعل الأمر للمثنى • أداة الاستفهام: أ (أسمعتَ؟) • باء القسم • القسم: أداة القسم والمقسم به وجواب القسم.	• المنادى	• مفردات تصف حالات مزاجية وشعورية إيجابية وسلبية مثل الفرح، السعادة، الغضب، الأسف، الحسرة	• هناك شيء من الرمزية على مستوى الحدث في الجملة يدركها القارئ بنفسه أو من خلال معارفه السابقة
ي	4	متوسط أوسط	≤15		• المبني للمجهول	• إن وأخواتها • كان وأخواتها • خبر مقدم / مبتدأ مؤخر • العنعنة/السند • رُبَّ (حرف جر شبيه بالزائد) • جملة الصلة وجملة الصفة • جملة الحال وجملة المفعول به	• أسماء الوصل المفردة • (قد – لقد) • (ممّا – عمّا – عمَّ – علامَ – فيمَ – إلامَ - بمَ...)	
ك		متوسط أعلى	≤20		• المشتقات العاملة (مثلا اسم الفاعل)	• جملة أسمية خبرها جملة أسمية • إضافة لفظية (طويل القامة)	• أسماء الوصل المثنى والجمع	• هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة
ل	5	متقدم أدنى			• التصغير	• جمل اعتراضية (تفسير، دعاء) • استثناء • حصر • بدل (مثلا بدل بعض أو اشتمال) • تمييز	• مفردات فصيحة - سامر III • اسم الفعل (مثلا آمين) • الأرقام (العربية أو الهندية > 1,000,000 • ذو • (بل - بلى - أجل - قط)	
م	6-7	متقدم أوسط			• نون التوكيد • تاء القسم	• الجمل شرطية (مركبة - عادية) • حرف الجزم لما	• كلمات تصف حالات نفسية عميقة مثل الاكتئاب، الضياع، الاستفار النفسي • استخدام كلمات منحوتة غير متداولة (مثلا هجرع للخفيف الأحمق مشتقة من هرع و هجع) • الرموز (ش.م.)	• أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث. • تعابير ثقافية محلية قد لا يفهمها من لا يشترك في نفس الثقافة.
ن	8-9	متقدم أعلى				• التوكيد المعنوي • المدح والذم • جملة أن المصدرية في محل رفع مبتدأ • صيغة التعجب "أفعل به من"	• مفردات فصيحة - سامر IV • مفردات قانونية، علمية، دينية، سياسية،... غير متخصصة/عامة • فو - حمو	
س	10-11	متقن أدنى				• تراكيب غير متداولة فيها التباس يحتاج إلى التشكيل الإعرابي لفكه	• المفردات المتخصصة التي لا تكفي معرفة الكلمة وحدها لفهمها، وإنما يحتاج إلى معرفة الفكرة/المفهوم لفهمها • الترخيم في أسماء العلم (مثلا أفاطم؟)	• أفكار رمزية، مجردة، علمية، أو شعرية وتحتاج إلى معارف لغوية ومعرفية سابقة للبناء عليها لأجل فهمها
ع	12	متقن أوسط					• مفردات فصيحة - سامر V • مفردات متخصصة ومفردات عربية عالية غير شائعة كثيرا في الفضاء العام. • مفردات في الغالب بعيدة عن اللهجات العامية.	
ف	جامعة 1-2	متقن أعلى					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص المبتدئ	
ص	جامعة 3-4	متفوق					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص	
ق	متخصص	متميز					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص الباحث	
هناك صعوبة	هذا الوسم يستخدم في حالة وجود صعوبة في تقييم المستوى، المفضل استخدام هذا الوسم حتى نتمكن كفريق عمل أن نجد حلا (مثلا بتعديل المعايير أو إضافة تفاصيل شرحية لها)							
هناك مشكلة	بصورة عامة، نستخدم هذا الوسم للجمل الحاوية على:				• أخطاء إملائية (مثلا همزات، تاء مربوطة، ألف مقصورة/ياء) • أخطاء في التشكيل • ركاكة لغوية (أمية، عامية، ترجمة سيئة من لغة أجنبية) • مواضيع غير لائقة (عنصرية، حيازية، تنمرية، إباحية، إلخ) • جمل وعبارات معظمها مكتوب بلغات غير العربية أو بغير الخط العربي	ولكن في الحالات التالية نوسم الجمل ونضيف أحد الحروف التالية في عامود الملاحظات: • خطأ في همزة الوصل/همزة القطع >> (أ) • كلمات خادشة >> (ع) • الخطأ في التشكيل في بداية الجملة >> (ت) • الياء غير المنقوطة في آخر الكلمة >> (ي)	

A.2 English Translation

BAREC Level	Grade	ACTFL	Word Count	Spelling/Pronunciation	Morphology	Syntax	Vocabulary	Idea/Content
1-alif	Pre1-1	Novice Low	1	• One-syllable and two-syllable words	• Singular imperfective verb	• One word	• Common noun • Proper noun (frequent and simple) • Personal pronouns (non-clitics) • Vocabulary identical to dialectal form - SAMER I • Numbers (Arabic or Indo-Arabic) 1-10	• Direct, explicit, and concrete idea. • No symbolism in the text.
2-ba	1	Novice Low	≤2	• Three-syllable words		• Nominal sentence • True idafa • Noun and adjective	• Verb • Adjective • Vocabulary similar to dialectal form - SAMER I • Spelled cardinal numbers • The five nouns: *Abw (father)*, *Axw (brother)*	
3-jim		Novice Mid	≤4		• Proclitic: Definite article *Al+* • Proclitic: Conjunction *wa+* • Enclitic: First Person Singular pronoun	• Apposition (full) • Demonstratives	• Common MSA vocabulary - SAMER I • Singular demonstrative pronoun • Numbers: 11-100	
4-dal		Novice Mid	≤6	• Words with an elongated Alif (e.g. /ʔāsif/)	• Plural imperfective verb • Prepositional proclitics • Nunated adverbials	• Verbal sentence w/o direct object • Preposition and object	• Prepositions	
5-ha	2	Novice High	≤8	• Four-syllable words	• Enclitic: Singular and Plural pronouns • Dual (in nouns and adjectives) • Sound feminine plural	• Verbal sentence with one nominal direct object • Conjoined sentences • Basic interrogative particles: what, when, who, where, how • Exclamatory form: how <comparative adjective>	• Ordinal numbers • Numbers: 101-1,000 • Dual and plural demonstrative pronoun	• Content is from the reader's life. • No symbolism in the text.
6-waw		Novice High	≤9	• Five-syllable words	• Singular and plural perfective verb • Sound masculine plural	• Sentence with two verbs (e.g., a verbal sentence a clausal direct object introduced with *Masdar 'an [~to/that]*)	• MSA vocabulary - SAMER I	
7-zay		Intermediate Low	≤10	• Six-syllable or more words • Verbs/nouns with weak final letters	• Dual perfective verb • Dual imperfective verb • Singular imperative verb • Enclitics: dual pronoun • Broken plurals • Waw of oath	• Adverbial accusative (time and place adverbs) • Circumstantial accusative • Interrogative particle *hal*	• High frequency MSA vocabulary - SAMER II	• Some symbolism, or not everything is stated directly in the sentence.
8-ha	3	Intermediate Low	≤11		• Plural imperative verb • Feminine plural suffix (*nun*) in nouns and verbs • Other proclitics: future *sa+*, continuation *wa+*, conjunction *fa+* • Conjunctions (e.g., then, until, or, whether, but, as for)	• Absolute object (emphasizing the verb) • Object of purpose • Object of accompaniment • Verbal sentence with two direct objects	• MSA vocabulary - SAMER I and II • Negation particles • Numbers: 1,001-1,000,000	• Some symbolism that requires the reader to seek help to understand the idea.
9-ta		Intermediate Mid	≤12		• Dual imperative verb • Interrogative Hamza • Ba of oath • Oath: The particle of oath, the object of the oath, and the answer to the oath	• Vocative	• Vocabulary describing positive and negative emotional and mood states like joy, happiness, anger, regret, sorrow	• Some symbolism at the event level in the sentence that the reader understands through prior knowledge.
10-ya	4	Intermediate Mid	≤15		• Passive voice	• *Inna* and its sisters (particles introducing a subject) • *Kana* and its sisters (past tense verbs) • Preposed predicate, postponed subject • Chain of narration • *rubba* preposition construction • Relative clauses • Circumstantial and object clauses	• Singular relative pronouns • Verbal particles *qad* and *laqad* • Preposition-Conjunctions: *mimma*, *fima*...	
11-kaf		Intermediate High	≤20		• Acting derivatives (e.g., the active participle)	• Nominal sentence with a nominal predicate • False idafa (tall in stature)	• Dual and plural relative pronouns	• A degree of symbolism and a need for prior knowledge to understand the meaning of the sentence.
12-lam	5	Advanced Low			• Diminutive form	• Parenthetical sentences (explanation, blessing) • Exception • Exclusivity • Apposition (e.g., partitive or containing) • Specification (*tamyiyz* construction)	• MSA vocabulary - Samer III • Frozen Verbs (e.g., *Āmiyn* Amen) • Numbers: > 1,000,000 • Five Nouns: Dhu (possession nominal) • Interjections: *bala*, *Ajal*, etc.	
13-mim	6-7	Advanced Mid			• Energetic mood (emphatic *nun*) • Ta of oath	• Conditional sentences (compound - simple) • Jussive particle *lamma* (not yet)	• Words describing deep psychological states like depression, loss, psychological alertness • Use of coined, uncommon words • Abbreviations (e.g., LLC)	• Symbolic ideas and deeper meanings, especially in terms of the psychological dimension of characters/events. • Local cultural expressions that may not be understood by those outside the culture.
14-nun	8-9	Advanced High				• Semantic emphasis • Praise and dispraise • *Masdar 'an* clause as a subject • Exclamatory form: <comparative adjective> *bih min*	• MSA vocabulary - SAMER IV • General legal, scientific, religious, political vocabulary, etc. • Five Nouns: *fw*, *Hmw*	
15-sin	10-11	Superior Low				• Uncommon constructions that are ambiguous and need diacritization for clarification	• Specialized vocabulary that requires understanding the concept/idea to comprehend it • Shortening in proper names (e.g., *fatim* for *fatima*)	• Symbolic, abstract, scientific, or poetic ideas that require prior linguistic and cognitive knowledge to understand.
16-ayn	12	Superior Mid					• MSA vocabulary - SAMER V • Specialized and highly elevated Arabic vocabulary not commonly used in public discourse. • Vocabulary mostly distant from dialects.	
17-fa	University Year 1-2	Superior High					• Scientific and heritage vocabulary not in use today, but familiar to a novice specialist	
18-sad	University Year 3-4	Distinguished					• Scientific and heritage vocabulary not in use today, but familiar to a specialist	
19-qaf	Specialist	Distinguished+					• Scientific and heritage vocabulary not in use today, but familiar to the advanced researcher specialist	
Difficulty	This tag is used when there is difficulty in assessing the level. It is preferred to use this tag so that the team can find a solution (for example, by adjusting the criteria or adding explanatory details).							
Problem	Generally, we use this tag for sentences containing:		• Spelling mistakes (e.g., Hamzas, Ta Marbuta, Alif maqsura/Ya) • Errors in diacritics • Linguistic awkwardness (illiteracy, colloquialism, poor translation from a foreign language) • Inappropriate topics (racism, bias, bullying, pornography, etc.) • Sentences and phrases mostly written in languages other than Arabic or in non-Arabic script		However, in the following cases, we provide the level and add a note in the comments column: • Error in Hamzat al-Wasl/Hamzat al-Qat' >> (ا) • Offensive words >> (ع) • Error in diacritics at the beginning of the sentence >> (ـّـ) • Dotted Yaa missing at the end of the word >> (ى)		

B Annotation Team

	A1	A2	A3
Native Language	Arabic	Arabic	Arabic
Other Language	En, Fr	En, Fr	En, Fr
Nationality	Lebanese	Lebanese	Lebanese
Residence	Lebanon	Lebanon	Lebanon
Gender	Female	Female	Female
Background	Muslim	Muslim	Muslim
Degree	BA	MA	B MA
Major	Geography	Arabic Lit.	Arabic Lit.
Experience	PT, LA	CT, LA	CT, LA, RA
School	-	Public&Private	Public
Level	Elementary	Secondary	Secondary
Students	L1	L1	L1
Years	22	22	25

Table 9: Annotator background information. All have extensive linguistic annotation experience. Certified Teacher (CT), Private Tutor (PT), Linguistic Annotator (LA), Research Assistant (RA).

C Annotation Batches Level Distribution

Level/Batch	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**
1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	4
2	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	0	12
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	34
4	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	1	1	10
5	2	3	2	3	2	3	2	3	2	3	3	3	3	3	3	3	3	3	49
6	3	3	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	55
7	7	7	7	6	8	7	7	7	8	7	8	8	8	8	8	8	8	7	134
8	7	7	7	7	7	7	7	7	7	7	7	7	8	8	8	8	7	7	130
9	3	3	3	3	4	4	3	3	4	3	3	3	3	3	3	3	5	5	61
10	11	12	11	12	11	12	12	12	11	12	12	12	11	12	11	12	11	12	209
11	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	89
12	24	24	24	24	23	24	23	23	23	23	23	23	24	24	24	23	24	23	423
13	5	5	6	5	5	5	6	5	5	5	6	5	5	5	5	5	5	5	93
14	19	19	19	18	19	18	18	19	19	19	18	19	19	18	19	19	19	19	337
15	4	4	4	5	5	4	5	4	4	4	4	4	4	4	4	4	4	4	75
16	3	2	2	3	2	3	2	3	3	3	3	2	3	2	2	3	3	3	47
17	1	2	1	2	1	1	1	1	1	1	1	2	1	2	1	1	1	2	23
18	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	3
19	0	0	0	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	5
Total	100	99	100	100	98	99	100	100	99	99	100	99	100	101	99	100	100	100	**1793**

Table 10: Distribution of sentences per annotation batch and readability level.

D Modeling Linguistic Phenomena

Feature extraction was conducted using a combination of CAMeL Tools (Obeid et al., 2020) , regular expressions, external lexicons, and custom Python scripts. Below is a breakdown of each feature group and the extraction methods used:

Morphological Features include three different categories:

- **Number of prefixes, suffixes, and clitics**: Extracted using CAMeL Tools' morphological disambiguator. Each token was decomposed into its base form and affixes, and counts were aggregated per sentence.
- **Verb tense and voice (e.g., passive, active)**: Identified using the POS tags and morphological features provided by CAMeL Tools.
- **Use of different forms (broken plurals, feminine plurals)**: Detected using morphological patterns and specific tag combinations (e.g.,singular form and plural num) from CAMeL analysis.

Table 12 shows the specific rules for all morphological features.

Syntactic Features A set of rule-based syntactic features was developed using dependency parsing outputs using the CamelParser (Elshabrawy et al., 2023). A dependency parse was first used to construct syntactic trees for each sentence, allowing for the identification of grammatical relations between words. From these structures, a set of binary features was extracted to reflect the presence or absence of key syntactic phenomena. Table 13 shows the specific rules for all Syntactic features.

Word Count Computed as the number of unique words in a sentence, ignoring repetitions, or punctuation.

Syllable Count The number of syllables in each word is computed by incorporating morphological and phonetic information. The CAPHI (consonant-vowel pattern) representation, the diacritized form of the word, and morphological prefix annotations are used for a more accurate count of syllables. The CAPHI string is tokenized and scanned for vowel segments, each indicating a potential syllable. Specific linguistic rules are applied to refine the syllable count:

- The final vowels are excluded if it is a diacritic (حركات الإعراب *Hrkat Al<ςrAb*).
- Morphological prefixes such as the definite article (ال التعريف *Al tςryf*) and conjunction (واو عاطفة *wAw ςATfp*) are excluded, as they do not contribute to the core syllabic structure of the main word.
- In the absence of CAPHI information, syllables are counted by identifying diacritic characters corresponding to short vowels within the diacritized form of the word.

Vocabulary-based Features The BAREC guidelines specify certain closed groups of vocabs that can be identified using the CAMeL Tools disambiguator, these are shown in Table 14.

E Linguistic codes and their associated readability levels

Tables 11, 12,13 14, and 15 present the complete set of linguistic features defined in the BAREC framework and used by the proposed rule-based model. Each feature is represented by a unique code and is associated with a specific readability level, reflecting the guideline followed by human annotators. Each code follows the format C*-n, where C denotes the linguistic category, * denotes the readability level, and n denotes the feature number within that category-level combination. The tables serve as a reference for the deterministic mapping between linguistic phenomena and readability levels employed by the system. Features whose Rule is '–' are defined in the guidelines but not implemented; the 67 implemented features are those with an extraction rule.

Feature	Feature (Arabic)	Rule	Code
Syllable count	عدد المقاطع	in D	O*-1
Words with an elongated Alif	كلمات تستخدم مد الألف	look for آ	O4-1
Verbs/nouns with weak final letters	أفعال / أسماء معتلة الآخر	pos=noun/verb, no suffix clitic, stem ends with ي ، ا ، or ى	O7-2
Word Count	عدد الكلمات	in D	W*-1

Table 11: Spelling, phonology and word-count features derived from the BAREC annotation guidelines.

Feature	Feature (Arabic)	Rule	Code
Singular imperfective verb	الفعل المضارع المفرد	num=s, asp=i, pos=verb	M1-1
Proclitic: Definite article Al+	سوابق: ال التعريف	prc0=Al_det	M3-1
Proclitic: Conjunction wa+	سوابق: واو العطف	prc2=wa_conj	M3-2
Enclitic: First person singular pronoun	لواحق: ضمير المتكلم المفرد المتصل	enc0=1s_pron / 1s_poss / 1s_dobj	M3-3
Plural imperfective verb	الفعل المضارع الجمع	pos=verb, asp=i, num=p	M4-1
Prepositional proclitics (bi+, li+, ka+)	سوابق: حروف جر متصلة (ب+، ل+، ك+)	prc1=bi_prep / li_prep / ka_prep	M4-2
Nunated adverbials	ظرف منون	-	M4-3
Enclitic: singular or plural pronoun	لواحق: ضمير متصل مفرد أو جمع	enc0 in [1p_dobj, . . . , 3p_pron]	M5-1
Dual (in nouns and adjectives)	المثنى (في الأسماء والصفات)	num=d, pos=noun / adj	M5-2
Sound feminine plural	جمع المؤنث السالم	form_num=p, form_gen=f, pos=noun / adj	M5-3
Singular and plural perfective verb	الفعل الماضي المفرد والجمع	pos=verb, asp=p, num=s / p	M6-1
Sound masculine plural	جمع مذكر سالم	form_gen=m, form_num=p, pos=noun / adj	M6-2

Continued on next page

Feature	Feature (Arabic)	Rule	Code
Dual perfective verb	الفعل الماضي المثنى	pos=verb, asp=p, num=d	M7-1
Dual imperfective verb	الفعل المضارع المثنى	pos=verb, asp=i, num=d	M7-2
Singular imperative verb	فعل الأمر المفرد	pos=verb, asp=c, num=s	M7-3
Enclitic: dual pronoun	لواحق: ضمير المثنى المتصل	enc0=[2d_dobj, . . . , 3d_pron]	M7-4
Broken plurals	جمع التكسير	pos=noun / adj, form_num=s, num=p	M7-5
Waw of oath (wa-Allāh)	واو القسم (والله)	prc2=wa_prep and lex ∈ qas-sam_lex	M7-6
Plural imperative verb	فعل الأمر الجمع	pos=verb, asp=c, num=p	M8-1
Feminine plural suffix (nun) in nouns and verbs	نون النسوة في الأسماء والأفعال	ends with noon and 3FP is present	M8-2
Other proclitics: future sa+, continuation wa+, conjunction fa+	سوابق أخرى: سين الاستقبال، واو الاستئناف، فاء العطف	prc0–prc3 = s_ / w_ / f_	M8-3
Conjunctions / connectives(e.g., then, until, or, whether, but, as for)	أدوات ربط (ثم ، حتى، أو، أم، لكن، أمّا)	lex ∈ connective_lex	M8-4
Dual imperative verb	فعل الأمر للمثنى	pos=verb, asp=c, num=d	M9-1
Interrogative Hamza	أداة الاستفهام: أ	prc3 in [>a_ques, A_ques]	M9-2
Ba of oath	باء القسم	prc1=bi_prep and lex ∈ qas-sam_lex	M9-3
Oath: the particle of oath, the object of the oath, and the answer to the oath	القسم : أداة القسم والمقسم به وجواب القسم.	-	M9-4
Passive voice	المبني للمجهول	vox=p, pos=verb	M10-1
Acting derivatives (e.g., the active participle)	المشتقات العاملة (مثلا اسم الفاعل)	-	M11-1
Diminutive	التصغير	-	M12-1

Continued on next page

Feature	Feature (Arabic)	Rule	Code
Energetic mood (emphatic nun)	نون التوكيد	-	M13-1
Ta of oath	تاء القسم	-	M13-2

Table 12: Morphological features and extraction rules derived from the BAREC annotation guidelines.

Feature	Feature (Arabic)	Rule	Code
One word	كلمة واحدة	one word	S1-1
True idafa	إضافة حقيقية	parent POS = NOM; pos=adj not in FEATS; has child with deprel IDF	S2-1
Nominal sentence	جملة اسمية	parent ≠ inna and sisters; has a child with dependency relation SBJ; POS ≠ VRB	S2-2
Noun + adjective	صفة وموصوف	adjacent NOM followed by ADJ (or noun with pos=adj); agreement in gender/number	S2-3
Full apposition (badal kull)	بدل كل	-	S3-1
Demonstrative apposition (badal ishāra)	بدل إشارة	-	S3-2
Verbal sentence without direct object	جملة فعلية بدون مفعول به	parent = verb; no dependency relation OBJ	S4-1
Preposition and object	جار ومجرور	parent POS: PRT; pos=prep in FEATS; has a child with dependency relation OBJ	S4-2
Verbal sentence with one nominal direct object	جملة فعلية مع مفعول به واحد اسم	parent = verb; has exactly one child with dependency relation OBJ	S5-1
Conjoined sentences	جمل معطوفة	-	S5-2
Basic interrogative particles: what, when, who, where, how	أدوات استفهام أساسية: ماذا، متى، من، أين، ما، كيف	lemma in {ماذا, متى, من, أين, ما, كيف}; POS indicates interrogative	S5-3
Exclamatory form: how <comparative adjective>	صيغة التعجب (ما أفعل)	-	S5-4
Sentence with two verbs (e.g. verbal sentence whose object is the masdar *an*)	جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	verb count ≥ 2; token lemma = أن; POS = PRT; has a child with POS = VRB, deprel = OBJ, asp=i	S6-1

Continued on next page

Adverbial accusative (time and place adverbs)	مفعول فيه (ظروف زمان ومكان)	-	S7-1
Circumstantial accusative (ḥāl)	حال	-	S7-2
Interrogative particle *hal*	أداة الاستفهام هل	lemma = هل; POS = `part_interrog`	S7-3
Absolute object (emphasizing the verb)	المفعول المطلق	-	S8-1
Object of purpose	المفعول لأجله	-	S8-2
Object of accompaniment	المفعول معه	-	S8-3
Verbal sentence with two direct objects	جملة فعلية تتعدى إلى مفعولين	parent POS = `VRB`; two children with dependency relation `OBJ`	S8-4
Vocative	المنادى	parent POS = `PRT`; FEATS include `pos=part_voc`; has child with deprel `OBJ`	S9-1
Inna and its sisters	إن وأخواتها	parent lemma in inna set; has a child with dependency relation `PRD`	S10-1
Kana and its sisters	كان وأخواتها	parent POS = `VRB`; lemma in kana set; has a child with dependency relation `PRD`	S10-2
Preposed predicate, postponed subject	خبر مقدم / مبتدأ مؤخر	dependency relation = `SBJ`; child POS $\neq$ `VRB`; parent index $<$ child index	S10-3
Chain of narration	العنعنة / السند	-	S10-4
Rubba preposition construction	رُبّ (حرف جر شبيه بالزائد)	-	S10-5
Relative clause, adjectival clause, ḥāl clause, object clause	جملة الصلة وجملة الصفة ـ جملة الحال ـ جملة المفعول به	-	S10-6
Nominal sentence whose predicate is a nominal sentence	جملة اسمية خبرها جملة اسمية	sentence does not start with a verb; contains child with deprel `TPC`	S11-1
False idafa (tall in stature)	إضافة خيالية (لفظية) طويل القامة	parent POS = `NOM`; `pos=adj` in FEATS; has child with deprel `IDF`	S11-2
Parentheticals (explanation, blessing)	جمل اعتراضية (تفسير ـ دعاء)	Arabic text inside parentheses, or in a non-initial/final comma segment	S12-1

Continued on next page

Exception (istithnā')	استثناء	POS = `part_restrict`	S12-2
Exclusivity	حصر	-	S12-3
Apposition (e.g., partitive or containing)	بدل (مثلا بدل بعض أو اشتمال)	-	S12-4
Specification (tamyīz)	تمييز	-	S12-5
Conditional sentences (compound, simple)	الجمل الشرطية (مركبة، عادية)	-	S13-1
Lammā as jussive particle	حرف الجزم لمّا	-	S13-2
Semantic emphasis (al-tawkīd al-ma'nawī)	التوكيد المعنوي	-	S14-1
Praise and dispraise	المدح والذم	-	S14-2
An-masdariyya clause as subject	جملة أن المصدرية في محل رفع مبتدأ	-	S14-3
Exclamatory form: <comparative adjective> bih min	صيغة التعجب (أفعل به من)	-	S14-4
Uncommon constructions that are ambiguous and need diacritization for clarification	تراكيب غير متداولة فيها التباس يحتاج إلى التشكيل الإعرابي لفكه	-	S15-1

Table 13: Syntactic features and extraction rules derived from the BAREC annotation guidelines.

Feature	**Feature (Arabic)**	**Rule**	**Code**
Proper noun (frequent and simple)	اسم علم (متداول بسيط تركيبيا)	pos = noun_prop	V1-1
Personal pronouns (non-clitics)	ضمير منفصل	pos = pron	V1-2
Common noun	اسم جنس	-	V1-3
Vocabulary identical to dialectal form – SAMER I	مفردات متطابقة مع العامية ـ سامر ١	-	V1-4

Continued on next page

Numbers (Arabic or Indo-Arabic) 1–10	الأرقام (العربية أو الهندية) ١-١٠	normalized numeric token where $1 \leq n \leq 10$	V1-5
The five nouns: Abw (father), Axw (brother)	الأسماء الخمسة: أبو، أخو	(lex = أَب and diac = أَبُو) or (lex = أَخ and diac = أَخُو), with pos = noun	V2-1
Vocabulary similar to dialectal form – SAMER I	مفردات متشابهة مع العامية ـ سامر ١	-	V2-2
Adjective	صفة	pos = adj	V2-3
Verb	فعل	pos = verb	V2-4
Spelled cardinal numbers	العدد الأصلي بالأحرف	pos = noun_num	V2-5
Singular demonstrative pronoun	اسم الإشارة المفرد	pos = pron_dem, num = s	V3-1
Common MSA vocabulary – SAMER I	مفردات فصيحة شائعة ـ سامر ١	-	V3-2
Numbers (Arabic or Indo-Arabic) 11–100	الأرقام (العربية أو الهندية) ١١-١٠٠	normalized numeric token where $11 \leq n \leq 100$	V3-3
Prepositions	حروف الجر	pos = prep, or prc1 ∈ {bi_prep, li_prep, ka_prep, wa_prep, ta_prep}	V4-1
Ordinal numbers	العدد الترتيبي	num = adj_num	V5-1
Numbers (Arabic or Indo-Arabic) 101–1,000	الأرقام (العربية أو الهندية) ١٠١-١٠٠٠	normalized numeric token where $101 \leq n \leq 1000$	V5-2
Dual and plural demonstrative pronouns	اسم اشارة مثنى، جمع	pos = pron_dem, num ∈ {d, p}	V5-3
MSA vocabulary – SAMER I	مفردات فصيحة ـ سامر ١	-	V6-1
Common MSA vocabulary – SAMER II	مفردات فصيحة شائعة ـ سامر ٢	-	V7-1
MSA vocabulary – SAMER I and SAMER II	مفردات فصيحة ـ سامر ١ و سامر ٢	-	V8-1

Continued on next page

Negation particles	أحرف النفي	pos == part_neg	V8-2
Numbers (Arabic or Indo-Arabic) 1,001–1,000,000	الأرقام (العربية أو الهندية) ١٠٠١-١٠٠٠٠٠٠	normalized numeric token where $1001 \leq n \leq 1{,}000{,}000$	V8-3
Vocabulary describing positive and negative moods and feelings (joy, happiness, anger, regret, sorrow)	مفردات تصف حالات مزاجية وشعورية إيجابية وسلبية مثل الفرح، السعادة، الغضب، الأسف، الحسرة	-	V9-1
Singular relative pronouns	أسماء الوصل المفردة	pos = pron_rel, num = s	V10-1
Verbal particles qad and laqad	(قد - لقد)	-	V10-2
Preposition-conjunctions: mimma, fima, …	(ممّا - عمّا - علامَ - فيمَ - إلامَ - بمَ...)	-	V10-3
Dual and plural relative pronouns	أسماء الوصل المثنى والجمع	pos = pron_rel, num ∈ {d, p}	V11-1
MSA vocabulary – SAMER III	مفردات فصيحة - سامر ٣	-	V12-1
Frozen verbs (e.g., Āmiyn Amen)	(اسم الفعل (مثلا آمين	-	V12-2
Numbers (Arabic or Indo-Arabic) > 1,000,000	الأرقام (العربية أو الهندية) 1,000,00 <	normalized numeric token where $n > 1{,}000{,}000$	V12-3
Five nouns: Dhu (possession nominal)	ذو	-	V12-4
Interjections: bal, bala, Ajal, qatt	(بل - بلى - أجل - قط)	-	V12-5
Words describing deep psychological states (depression, loss, psychological mobilization)	كلمات تصف حالات نفسية عميقة مثل الاكتئاب، الضياع، الاستنفار النفسي	-	V13-1
Uncommon coined words	استخدام كلمات منحوتة غير متداولة	-	V13-2
Symbols and abbreviations (e.g., LLC)	(الرموز (شم.	pos = abbrev	V13-3

Continued on next page

MSA vocabulary – SAMER IV	مفردات فصيحة ـ سامر ٤	-	V14-1
Legal, scientific, religious, political (etc.) non-specialized/general vocabulary	مفردات قانونية، علمية، دينية، سياسية،... غير متخصصة / عامة	-	V14-2
Five nouns: *fw*, *Hmw*	فو ـ حمو	-	V14-3
Specialized vocabulary where knowing the word alone is insufficient; the concept must be understood	المفردات المتخصصة التي لا تكفي معرفة الكلمة وحدها لفهمها، وإنما يحتاج إلى معرفة الفكرة لفهمها	-	V15-1
Shortening in proper names (e.g., fatim for fatima)	الترخيم في أسماء العلم : مثلا أفاطم؟	-	V15-2
MSA vocabulary – SAMER V	مفردات فصيحة ـ سامر ٥	-	V16-1
Specialized and high-register Arabic vocabulary uncommon in public discourse; mostly removed from colloquial dialects	مفردات متخصصة ومفردات عربية عالية غير شائعة كثيرا في الفضاء العام. مفردات في الغالب بعيدة عن اللهجات العامية.	-	V16-2
Scientific and heritage vocabulary not in use today, but familiar to a novice specialist	مفردات عِلمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص المبتدئ	-	V17-1
Scientific and heritage vocabulary not in current use; familiar to a specialist	مفردات عِلمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص	-	V18-1
Scientific and heritage vocabulary not in current use; familiar to the advanced researcher specialist	مفردات عِلمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص الباحث	-	V19-1

Table 14: Vocabulary features derived from the BAREC annotation guidelines.

Feature	Feature (Arabic)	Rule	Code
Direct, explicit, and concrete idea; no symbolism in the text	فكرة مباشرة وصريحة وحسية. لا رمزية في النص.	-	C1-1
Content drawn from the reader's life; no symbolism in the text	المحتوى من حياة القارئ. لا رمزية في النص.	-	C5-1
Some symbolism, or not everything is stated directly in the sentence	بعض الرمزية أو عدم التصريح المباشر بكل المقصود في الجملة	-	C7-1
Some symbolism for which the reader needs assistance to grasp the intended meaning	بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة	-	C8-1
Some symbolism at the event level of the sentence that the reader can grasp alone or through prior knowledge	هناك شيء من الرمزية على مستوى الحدث في الجملة يدركها القارئ بنفسه أو من خلال معارفه السابقة	-	C9-1
A degree of symbolism and need for prior knowledge to understand the intended meaning	هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة	-	C11-1
Symbolic and underlying meanings, especially about the psychological dimension of characters or events; local cultural expressions that may not be understood by those outside the culture	أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث. تعابير ثقافية محلية قد لا يفهمها من لا يشترك في نفس الثقافة.	-	C13-1
Symbolic, abstract, scientific, or poetic ideas that require prior linguistic and cognitive knowledge to build on for understanding	أفكار رمزية، مجردة، عِلمية، أو شعرية وتحتاج إلى معارف لغوية ومعرفية سابقة للبناء عليها لأجل فهمها	-	C15-1

Table 15: Content/conceptual features derived from the BAREC annotation guidelines.

F LLM Prompt Template

System message:

You are an expert Arabic linguist.

You will be given a sentence and a list of linguistic features (each with a code and description). Identify which features are present in the sentence. Return ONLY a valid JSON array of the matching codes, with no explanation. If none match, return an empty array: `[]`

Available features:

– Code: W8-1 | Feature: ١١ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم

– Code: M8-1 | Feature: فعل الأمر الجمع

– Code: M8-2 | Feature: نون النسوة في الأسماء والأفعال (انتظرنَ دورهنّ)

– Code: M8-3 | Feature: سوابق أخرى: سين الاستقبال ، واو الاستئناف، فاء العطف

– Code: M8-4 | Feature: (ثم ، حتى ، أو ، أم ، لكن ، أمّا)

– Code: S8-1 | Feature: المفعول المطلق

– Code: S8-2 | Feature: المفعول لأجله

– Code: S8-3 | Feature: المفعول معه

– Code: S8-4 | Feature: جملة فعلية تتعدى إلى مفعولين

– Code: V8-1 | Feature: مفردات فصيحة ـ سامر ١ و سامر ٢

– Code: V8-2 | Feature: أحرف النفي

– Code: V8-3 | Feature: الأرقام (العربية أو الهندية) ١,٠٠١-١,٠٠٠,٠٠٠

– Code: C8-1 | Feature: بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة

Important counting instructions:

– For word count (WC) features: carefully count the number of unique non-repeating typographic words in the sentence, strictly excluding all punctuation marks. Only mark WC if the unique word count exactly matches the number described in the feature.

User message:

Here are some annotated examples to guide you:

[Example 1]

Sentence: والعنكبوت تفضله بعض مناطق أمريكا الجنوبية!

Answer: `["M8-3"]`

[Examples 2 through k follow the same format, drawn from the opposite cross-validation fold.]

Now it's your turn.

Sentence to analyse:

واحصد الخيرات من الثمار

Figure 2: Few-shot prompt template for level 8 (*ha*). The system message lists the BAREC features associated with the target level. The user message contains k annotated exemplars, followed by the target sentence. For $k = 0$, the exemplar block is omitted and the user message contains only the target sentence.

G Per-Level Results

	Exact			Jaccard			At-Least-1			Precision			Recall			F1		
Level	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**
1	25.0	**50.0**	25.0	**77.8**	75.0	71.4	100.0	100.0	100.0	77.8	85.7	**100.0**	**100.0**	85.7	71.4	**87.5**	85.7	83.3
2	25.0	41.7	**50.0**	25.0	37.5	**50.0**	50.0	50.0	**66.7**	27.3	42.9	**60.0**	75.0	75.0	75.0	40.0	54.5	**66.7**
3	55.9	**58.8**	38.2	67.3	**70.0**	58.3	94.1	94.1	**97.1**	81.4	**85.4**	68.6	79.5	79.5	79.5	80.5	**82.4**	73.7
4	60.0	60.0	60.0	**70.0**	**70.0**	66.7	90.0	90.0	90.0	**73.7**	**73.7**	70.0	93.3	93.3	93.3	**82.4**	**82.4**	80.0
5	18.4	**49.0**	16.3	42.1	**58.3**	25.7	**93.9**	89.8	36.7	43.2	**62.0**	52.9	**94.4**	90.7	33.3	59.3	**73.7**	40.9
6	49.1	**60.0**	50.9	46.2	**51.9**	50.7	**72.7**	**72.7**	63.6	56.0	64.6	**76.1**	**72.4**	**72.4**	60.3	63.2	**68.3**	67.3
7	45.5	**70.1**	47.8	58.7	**71.1**	52.5	**92.5**	88.8	61.2	62.5	80.5	**86.7**	**90.6**	85.9	57.1	74.0	**83.1**	68.8
8	63.1	**63.8**	59.2	69.1	**70.1**	62.5	**87.7**	86.9	72.3	78.6	81.9	**88.5**	**85.0**	83.0	68.0	81.7	**82.4**	76.9
9	80.3	**88.5**	50.8	82.7	**89.7**	51.5	**96.7**	95.1	55.7	87.3	96.8	**100.0**	**93.9**	92.4	51.5	90.5	**94.6**	68.0
10	**72.7**	68.9	51.7	**75.6**	73.5	56.2	**90.0**	86.6	73.2	**83.5**	83.1	74.4	**88.9**	86.3	69.7	**86.1**	84.7	72.0
11	39.3	13.5	**47.2**	40.3	15.5	**46.5**	**64.0**	27.0	52.8	52.3	27.3	**82.5**	**63.7**	26.4	51.6	57.4	26.8	**63.5**
12	57.9	48.2	**70.0**	62.8	49.6	**64.8**	**88.4**	64.3	83.9	73.5	76.8	**83.8**	**81.1**	58.3	74.1	77.1	66.3	**78.6**
13	58.1	**59.1**	**59.1**	59.1	**59.8**	43.5	**82.8**	77.4	61.3	67.8	**73.0**	61.3	**82.1**	76.8	60.0	74.3	**74.9**	60.6
14	93.5	92.6	**93.8**	**93.4**	92.3	91.6	**98.8**	98.5	96.7	95.5	94.9	**96.7**	**97.7**	97.1	94.5	**96.6**	96.0	95.6
15	45.3	41.3	**77.3**	58.2	48.0	**76.1**	89.3	78.7	**93.3**	67.0	62.1	**93.3**	**81.6**	67.8	80.5	73.6	64.8	**86.4**
16	**95.7**	93.6	**95.7**	**95.9**	93.9	**95.9**	**100.0**	97.9	**100.0**	100.0	100.0	100.0	**95.9**	93.9	**95.9**	**97.9**	96.8	**97.9**
17	82.6	**87.0**	**87.0**	84.0	**88.0**	84.6	91.3	**95.7**	**95.7**	**100.0**	**100.0**	95.7	84.0	**88.0**	**88.0**	91.3	**93.6**	91.7
18	66.7	66.7	**100.0**	66.7	66.7	**100.0**	66.7	66.7	**100.0**	100.0	100.0	100.0	66.7	66.7	**100.0**	80.0	80.0	**100.0**
19	80.0	80.0	80.0	83.3	83.3	83.3	100.0	100.0	100.0	100.0	100.0	100.0	83.3	83.3	83.3	90.9	90.9	90.9

Table 16: Per-level results (%) under the Category-only (C) setting for Gemini 3 (Gem), GPT-5.5 (GPT), both 4-shot, and the rule-based system (Rules). Precision, recall, and F1 are micro-averaged within each level.

	Exact			Jaccard			At-Least-1			Precision			Recall			F1		
Level	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**	**Gem**	**GPT**	**Rules**
1	0.0	**50.0**	25.0	65.0	**75.0**	71.4	100.0	100.0	100.0	68.4	85.7	**100.0**	**92.9**	85.7	71.4	78.8	**85.7**	83.3
2	16.7	25.0	**33.3**	18.5	26.3	**26.7**	41.7	41.7	**50.0**	21.7	33.3	**40.0**	**55.6**	**55.6**	44.4	31.3	41.7	**42.1**
3	55.9	**58.8**	35.3	66.1	**68.5**	57.8	94.1	94.1	**97.1**	78.7	**82.2**	67.3	80.4	80.4	80.4	79.6	**81.3**	73.3
4	**50.0**	**50.0**	40.0	**63.6**	60.9	50.0	**90.0**	**90.0**	80.0	**66.7**	63.6	54.2	**93.3**	**93.3**	86.7	**77.8**	75.7	66.7
5	18.4	**49.0**	16.3	42.1	**58.3**	25.7	**93.9**	89.8	36.7	43.2	**62.0**	52.9	**94.4**	90.7	33.3	59.3	**73.7**	40.9
6	49.1	**60.0**	50.9	46.2	**51.9**	50.7	**72.7**	**72.7**	63.6	56.0	64.6	**76.1**	**72.4**	**72.4**	60.3	63.2	**68.3**	67.3
7	41.0	**64.9**	46.3	55.0	**65.1**	48.8	**91.8**	87.3	59.7	58.8	74.4	**81.2**	**89.3**	83.9	55.0	70.9	**78.9**	65.6
8	60.0	**60.8**	53.8	67.4	**68.7**	58.6	**87.7**	86.9	71.5	76.2	79.6	**83.9**	**85.3**	83.3	66.0	80.5	**81.4**	73.9
9	80.3	**88.5**	50.8	82.7	**89.7**	51.5	**96.7**	95.1	55.7	87.3	96.8	**100.0**	**93.9**	92.4	51.5	90.5	**94.6**	68.0
10	48.8	**50.2**	45.5	61.1	**61.8**	49.4	**87.1**	86.1	68.4	67.4	69.0	**69.3**	**86.8**	85.6	63.2	75.9	**76.4**	66.1
11	39.3	13.5	**44.9**	38.1	13.2	**43.3**	**61.8**	23.6	50.6	50.5	23.9	**78.9**	**60.9**	22.8	48.9	55.2	23.3	**60.4**
12	52.2	42.8	**59.8**	**57.2**	45.7	53.1	**86.8**	62.6	75.4	66.7	70.6	**74.9**	**80.0**	56.5	64.6	**72.7**	62.8	69.3
13	52.7	**53.8**	2.2	49.6	**52.7**	21.1	**74.2**	72.0	51.6	60.3	**66.7**	26.5	**73.7**	71.6	50.5	66.4	**69.0**	34.8
14	76.6	**78.3**	9.8	77.8	**78.2**	51.9	96.4	95.3	**96.7**	86.4	**89.6**	53.4	88.7	86.0	**95.0**	87.5	**87.8**	68.4
15	45.3	41.3	**77.3**	58.2	48.0	**76.1**	89.3	78.7	**93.3**	67.0	62.1	**93.3**	**81.6**	67.8	80.5	73.6	64.8	**86.4**
16	70.2	**76.6**	10.6	**73.8**	73.3	54.2	97.9	91.5	**100.0**	81.4	**88.0**	55.3	88.9	81.5	**96.3**	**85.0**	84.6	70.3
17	82.6	**87.0**	**87.0**	84.0	**88.0**	84.6	91.3	**95.7**	**95.7**	**100.0**	**100.0**	95.7	84.0	**88.0**	**88.0**	91.3	**93.6**	91.7
18	66.7	66.7	**100.0**	66.7	66.7	**100.0**	66.7	66.7	**100.0**	100.0	100.0	100.0	66.7	66.7	**100.0**	80.0	80.0	**100.0**
19	80.0	80.0	80.0	83.3	83.3	83.3	100.0	100.0	100.0	100.0	100.0	100.0	83.3	83.3	83.3	90.9	90.9	90.9

Table 17: Per-level results (%) under the Category+Reason (CR) setting for Gemini 3 (Gem), GPT-5.5 (GPT), both 4-shot, and the rule-based system (Rules). Precision, recall, and F1 are micro-averaged within each level.

H License

We list below the licenses of the data, tools, and models used in this work, all of which are employed in accordance with their intended use.

- SAMER Lexicon (Al Khalil et al., 2020): Non Commercial Use - NYUAD License.[6]
- BAREC Corpus (Elmadani et al., 2025): Creative Commons Attribution Share Alike 4.0
- CAMeL Tools (Obeid et al., 2020) and CAMeL-BERT (Inoue et al., 2021): MIT License
- CamelParser2.0 (Elshabrawy et al., 2023): MIT License
- *ALLaM-7B-Instruct* (Bari et al., 2025): Apache license 2.0
- *Fanar-2-27B-Instruct* (TEAM et al., 2026): Apache license 2.0
- *Jais-2-70B-Chat* (Anwar et al., 2025): Apache license 2.0
- *Qwen3-30B-A3B-Instruct* (Team, 2025): Apache license 2.0
- *Aya-Expanse-32B* (Dang et al., 2024): Creative Commons Attribution Non Commercial 4.0
- *Llama-3.3-70B-Instruct* (Grattafiori et al., 2024): LLAMA 3.3 COMMUNITY LICENSE AGREEMENT[7]

[6] https://camel.abudhabi.nyu.edu/samer-readability-lexicon/

[7] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct/blob/main/LICENSE