A Bilingual Bimodal Benchmark for Arabic-English NLP Across Grammatical Correction, Essay Scoring, Morphological Tagging, and Speech Recognition

Bashar Alhafni[1] **Injy Hamed**[1] **Fadhl Eryani**[2,4]
David Palfreyman[3] **Nizar Habash**[1,4]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]University of Tübengen, [3]United Arab Emirates University
[4]New York University Abu Dhabi
{bashar.alhafni,injy.hamed}@mbzuai.ac.ae, dpalf@uaeu.ac.ae
{fadhl.eryani,nizar.habash}@nyu.edu

Abstract

Building comprehensive datasets that support a variety of NLP tasks and cover a diversity of languages and domains is vital for NLP evaluation purposes. In this paper, we present **ZAEBUC***, a dataset that builds upon and enriches prior corpora with new annotations and benchmarking experiments. ZAEBUC* serves as a benchmark for a range of NLP tasks, including grammatical error correction, automated essay scoring, automatic speech recognition, and morphological tagging, which includes tokenization, part-of-speech tagging, and lemmatization. The dataset covers Arabic and English in both written and spoken forms, offering a bilingual and bimodal resource. Furthermore, the corpus brings together a collection of resources gathered from a similar population, enabling cross-linguistic and cross-modal comparisons. We provide benchmarking results, demonstrating the performance of NLP models, including LLMs, across various tasks, languages, and modalities.

Keywords: Arabic, English, Grammatical Correction, Essay Scoring, Morphological Tagging, Speech Recognition

1. Introduction

In computational linguistics, developing comprehensive datasets for diverse natural language processing (NLP) tasks is essential, but many existing datasets are limited in scope. For example, morphological annotations are often separate from those for tasks like grammatical error correction (GEC), automated essay scoring (AES), or automatic speech recognition (ASR). Moreover, language corpora tend to focus on a specific language rather than gathering data from bilingual writers, overlooking the unique research questions that could be answered from relating native language (L1) with second-language (L2) writing at various linguistic levels, including spelling, vocabulary, and language proficiency. Ultimately, such datasets should be collected from a similar population of users, allowing for more consistent control of variables and leading to more meaningful comparisons.

To bridge this gap and inspired by the Zayed University Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) Project (Habash and Palfreyman, 2022; Hamed et al., 2024) that comprises datasets collected from undergraduate students at Zayed University in the UAE covering different NLP tasks, we extend their efforts by presenting the **ZAEBUC*** corpus, which consolidates all previous ZAEBUC efforts into a unified resource. ZAEBUC* encompasses multiple word-level, sentence-level, and document-level annotations that enable multiple NLP tasks such as GEC, AES, ASR, and morphological tagging, including tokenization, part-of-speech (POS) tagging, and lemmatization. Our corpus spans both Arabic and English, in both written and spoken forms. We demonstrate the usability of our corpus by presenting benchmarking results across different tasks, languages, and modalities. By creating ZAEBUC*, we aim to introduce a unified resource for advancing cross-linguistic and cross-modal research in Arabic and English NLP. We make ZAEBUC* publicly available to support and encourage further research and development in these areas.[1]

2. Related Work

The growing availability of text corpora, varying in size, genre, and types of annotation, has significantly advanced NLP research across different tasks. This includes GEC (Ng et al., 2014; Mohit et al., 2014; Rozovskaya et al., 2015; Napoles et al., 2017; Bryant et al., 2019; Náplava et al., 2022, *inter alia*), AES (Yannakoudakis et al., 2011; Boyd et al., 2014; Mathias and Bhattacharyya, 2018; Bashendy et al., 2024, 2026, *inter alia*), morphosyntactic tagging (Marcus et al., 1993; Nivre et al., 2016; Habash et al., 2022, *inter alia*), and ASR (Ali et al., 2016,

[1] http://www.zaebuc.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1732–1749
11–16 May 2026.

2017; Mubarak et al., 2021; Hamed et al., 2022, *inter alia*). However, despite the interrelatedness of these tasks, many available datasets are collected independently, often in isolation from one another. This separation limits the potential for cross-task learning, as these tasks could greatly benefit from shared data resources and annotations. For example, AES could benefit from GEC in assessing language proficiency (Doi et al., 2024), while morphosyntactic tagging could support both by providing richer linguistic context (Vajjala and Rama, 2018; Alhafni et al., 2023; Li and Ng, 2024). Similarly, ASR could benefit from incorporating GEC, enhancing both speech recognition and language understanding (D'Haro and Banchs, 2016; Tanaka et al., 2018). Moreover, many dataset creation efforts focus on a single language and overlook bilingual users. This monolingual emphasis limits the study of language switching, cross-linguistic transfer from a native language (L1) to a second language (L2), and proficiency development in multilingual contexts (Cook, 2016).

To address the above limitations in existing datasets, the Zayed University Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) Project was established. ZAEBUC focuses on bilingual users of Arabic and English, and comprises samples of their writing and speech in both languages. The written corpus of ZAEBUC was introduced by Habash and Palfreyman (2022) and it consists of Arabic and English annotated essays written by 397 first-year university students at Zayed University in the UAE. The corpus comprises 388 English essays (97.5K words) and 214 Arabic essays (34K words). The corpus annotations include: (1) anonymized metadata indicating extralinguistic features of the writers and texts; (2) manually spelling and grammar corrected versions of the raw text; (3) manual morphological tagging, including tokenization, POS, and lemmatization; and (4) writing proficiency ratings using the Common European Framework of Reference (CEFR) (Council of Europe, 2001).

Hamed et al. (2024) introduced ZAEBUC-Spoken, a dataset comprising twelve hours of Zoom meetings in which multiple speakers engage in a role-playing work scenario. In these meetings, students from Zayed University brainstorm ideas on a given topic and then discuss them with an interlocutor. The corpus is multilingual, featuring (accented) English, Modern Standard Arabic (MSA), and two Arabic dialects–Gulf and Egyptian–spoken by individuals from six different nationalities. Also, given the prevalence of code-switching in the Arab world (Hamed et al., 2025), the speakers frequently code-switch between the four languages. The dataset includes manual transcriptions of the recordings, dialectness level annotations for segments where code-switching occurs between Arabic varieties, and automatic morphological annotations.

While the ZAEBUC project provides valuable written and spoken data, a comprehensive evaluation of NLP models across its tasks is still lacking. Prior work using ZAEBUC has focused only on GEC (Alhafni et al., 2023; Alhafni and Habash, 2025) and AES (Qwaider et al., 2025a,b).

In this work, we build on previous ZAEBUC efforts and extend the written corpus of ZAEBUC by collecting additional Arabic and English essays. The essays are manually annotated with grammar and spelling corrections, CEFR labels, and morphological tags, including tokens, POS, and lemmas. We further enrich ZAEBUC-Spoken with manual morphological annotations. Moreover, we provide benchmarking results across the various tasks supported by ZAEBUC including, GEC, AES, morphological tagging, and ASR, for Arabic and English.

3. Datasets

In this section, we discuss the extensions to previously collected ZAEBUC datasets in both written and spoken domains. In the written domain, we provide a new dataset that follows the same data collection and annotation processes as Habash and Palfreyman (2022). Throughout the paper, we refer to the initial data collected by Habash and Palfreyman (2022) as ZAEBUC-W1 and our newly collected dataset as ZAEBUC-W2. In the spoken domain, we enrich the ZAEBUC-Spoken dataset collected by Hamed et al. (2024) with manual morphological annotations. Table 1 presents an overview on the current state of the three datasets, outlining the size of the corpora as well as the present annotations. In Tables 2 and 3, we provide examples, showing the annotations provided in the ZAEBUC corpora.

3.1. ZAEBUC-Written

We provide a summary of the annotation decisions and statistics for the ZAEBUC-Written corpora.

Corpus Collection Similar to the process used for the ZAEBUC-W1 corpus, we obtain approval from Zayed University's Research Ethics Committee to collect data for ZAEBUC-W2. The student writers in the ZAEBUC-W1 corpus were in their first semester, enrolled in two Freshman writing courses–one in Arabic and one in English. For ZAEBUC-W2, two years later, we identified a set of fifth-semester core courses across various university majors, aiming to include some of the same students from ZAEBUC-W1. These courses also included other students from earlier and later cohorts. We coordinated with faculty teaching these courses to have students write on the same topics as those

	ZAEBUC-W1		ZAEBUC-W2		ZAEBUC-S		
	AR	**EN**	**AR**	**EN**	**AR**	**EN**	**CSW**
# Docs / Utts.	214	388	104	308	1,835	2,491	930
# Tokens	34,235	97,478	21,358	70,009	23,326	48,558	21,257
GEC	✔		✔*		✗		
CEFR	✔		✔*		✗		
Morph	✔		✔*		✔*		
Speech Transcriptions	✗		✗		✔		

Table 1: Summary of statistics and annotation tasks in the ZAEBUC-Written (W1 and W2) and ZAEBUC-Spoken (S) Corpora across Arabic (AR), English (EN), and Code-Switching (CSW). We report the number of documents (# Docs) for ZAEBUC-Written and number of utterances for ZAEBUC-Spoken (# Utts). ✔denotes applicability, while ✗denotes non-applicability. * denotes newly added data and annotations.

	Arabic	**English**
Input	كما نعلم ان التواصل الاجتماعي له تأثيرعلى الفرد و المجتمع و له دور كبير على الاقتصاد و الثقافه و العادات و التقاليد ايضا . التواصل الاجتماعي له سلبيات و ايجابيات و من إجابياته تسهيل التواصل مع الاخرين , زيادة الوعي بالقضايا المجتمعيه , تعزيز القوة الاقتصاديه و الثقافيه , اما عن الاثار السلبيه هي الحد المباشر من التواصل بين افراد المجتمع , نشر الاشاعات و الاخبار الكاذبه , انتهاك الخصوصيه و غيره الكثير مثل الجرائم الالكترونيه . اصبح التواصل الاجتماعي جزء لا يتجزأ من حياة الكثيرمنا بعد انتشارة السريع على مدى السنين , ولا يقتصر استخدامه على فئة عمريه معينه بل انه متاح لي جميع الفئات العمريه , وايضا هي وسيله اجتماعيه ممتعه لا تقف عند حد معين .	social media has a lot of positivities and negativities on the indvidual. i think most negativities are happening to people that are young or the people that cant control the use of the internet. the internet is not life, not everything. social media helps me study, communicate, learn more, source of knowlage. to me social media is a positive place and the main reason is the control i have over it not it controling me. when you can control the use of social media it becomes better and the older the better you know how to use it like this is right to post, say, use, see. but kids dont know that and young people are out of control.
GEC	كما نعلم أن التواصل الاجتماعي له تأثير على الفرد والمجتمع وله دور كبير على الاقتصاد والثقافة والعادات والتقاليد أيضا . التواصل الاجتماعي له سلبيات وإيجابيات ومن إيجابياته تسهيل التواصل مع الآخرين , وزيادة الوعي بالقضايا المجتمعية , وتعزيز القوة الاقتصادية والثقافية , أما عن الآثار السلبية هي الحد المباشر من التواصل بين أفراد المجتمع , ونشر الإشاعات والأخبار الكاذبة , وانتهاك الخصوصية وغيرها الكثير مثل الجرائم الإلكترونية . أصبح التواصل الاجتماعي جزءا لا يتجزأ من حياة الكثير منا بعد انتشاره السريع على مدى السنين , ولا يقتصر استخدامه على فئة عمرية معينة بل إنه متاح لجميع الفئات العمرية , وأيضا هو وسيلة اجتماعية ممتعة لا تقف عند حد معين .	Social media has a lot of positives and negatives for the individual. I think most negatives are happening to people who are young or the people who can't control the use of the Internet. The Internet is not life nor everything. Social media helps me study, communicate and learn more, as a source of knowledge. For me, social media is a positive place, and the main reason is the control I have over it, not it controlling me. When you can control the use of social media, it becomes better, and the older you are, the better you know how to use it; like, this is right to post, say, to use or to see. But kids don't know that, and young people are out of control.
CEFR	**Label 1:** B1 **Label 2:** B2 **Label 3:** B1 **Avg Label:** B1	**Label 1:** B1 **Label 2:** B1 **Label 3:** A2 **Avg Label:** B1
Morph	**Tokens** كما نعلم أن التواصل الاجتماعي ل+ه تأثير على الفرد و+المجتمع و+ل+ه دور كبير على الاقتصاد...	**Tokens** Social media has a lot of positives and negatives for the individual …
	Lemmas كَما عَلِم أَنَّ تَواصُل اجْتِماعِيّ لِ تَأْثِير عَلَى فَرْد مُجْتَمَع لِ دَوْر كَبِير عَلَى اقْتِصاد ...	**Lemmas** social media have a lot of positive and negative for the individual …
	POS CCONJ VERB SCONJ NOUN ADJ ADP+PRON NOUN ADP NOUN …	**POS** ADJ NOUN VERB DET NOUN ADP NOUN CCONJ NOUN ADP DET NOUN …

Table 2: Arabic and English examples from the ZAEBUC-Written corpus: original text (Input), spelling and grammar correction (GEC), CEFR levels, and morphological annotations (tokenization, lemmatization, and POS). Red words indicate errors; green words indicate corrections.

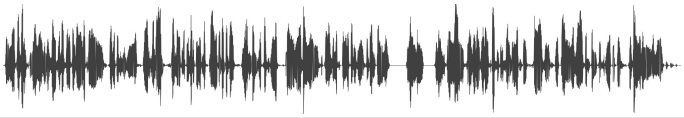
Audio		
Transcription	هيه ويفضل يكون ال-- و-- ويفضل فريق العمل يكونون من فريق الش-- من من الشباب عشان تعطى يعطونج أفكار متجددة و-- .. للتسويق، لأنهم هم دايما في التواصل الإجتماعي فعندهم هالخبرة القوية. وال-- وال+online فيه عندهم online courses يساعدونج في إنه ي-- يقوون مهاراتج في التسويق وفي في جميع المهارات.	
Morph	**Tokens**	هيه و+يفضل يكون ال-- و-- و+يفضل فريق العمل يكونون من فريق الش-- من من الشباب عشان ...
	Lemmas	هَيْه فَضَّل كان ال-- و-- فَضَّل فَريق عَمَل كان مِنْ فَريق الش-- مِنْ مِنْ شابّ عَشان ...
	POS	VERB CCONJ+VERB VERB <PW> <PW> CCONJ+VERB NOUN NOUN VERB ADP NOUN <PW> ADP ADP NOUN SCONJ ...

Table 3: Example from the ZAEBUC-Spoken corpus: besides the audio, we have the transcription and morphological annotations (tokenization, lemmatization, and POS).

in ZAEBUC-W1 (social media, national development, and tolerance). All participating students provided written consent to release their work.

While students in ZAEBUC-W1 had submitted their writing through a secure browser with proofing tools disabled, this approach was not feasible for all students in ZAEBUC-W2. Therefore, we allowed submissions via Google Forms, which gave students access to spelling and grammar checking tools, as well as the ability to consult online sources. A plagiarism check revealed that less than 4% of the Arabic texts and 11% of the English texts in ZAEBUC-W2 contained a significant proportion (20% or more) of plagiarized content.

CEFR Annotation The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) categorizes language learners' abilities in speaking, reading, listening, and writing. It uses six proficiency levels, ranging from A1 (Beginner) to C2 (Proficient), to classify users by language skills. In our annotation process, each essay was independently annotated by three CEFR-proficient bilingual speakers (fluent in Arabic and English). The annotators assigned a CEFR level to each essay and provided comments supporting their assessments. Following Habash and Palfreyman (2022), we assign a holistic CEFR level to each essay by converting the three CEFR ratings into numerical scores (ranging from 1 to 6) and then taking the rounded average. If an essay is considered to be unassessable by one annotator, then the average will be unassessable. When it comes to the inter-annotator agreement (IAA), the average pairwise exact agreement among the three annotators is 61% for Arabic and 62% for English.

	ZAEBUC-W1		ZAEBUC-W2	
	AR	**EN**	**AR**	**EN**
A1	0	2%	0	0
A2	3%	24%	0	4%
B1	51%	50%	49%	61%
B2	37%	21%	45%	31%
C1	5%	3%	4%	3%
C2	0	0	0	0
Unassessable	3%	0	2%	2%

Table 4: CEFR levels distributions in the ZAEBUC-Written Corpora.

The average pairwise Quadratic Weighted Cohen's Kappa (Cohen, 1968) is 0.56 for Arabic (moderate agreement) and 0.43 for English (fair agreement). Table 4 presents the distributions of the average CEFR levels for both Arabic and English essays. Most Arabic essays fall within the B1–B2 range despite being written by native speakers, reflecting the diglossic nature of Arabic and the fact that MSA is primarily acquired through formal instruction.

Text Correction For spelling and grammar correction, we followed the same guidelines used by Habash and Palfreyman (2022), which are inspired by Zaghouani et al. (2014) for Arabic and Dahlmeier et al. (2013) for English. The annotators were instructed to correct various spelling and grammar errors (e.g., morphological and orthographic errors), while avoiding making changes to the lexical choices made by the writers except for closed-class terms such as prepositions, pronouns and articles. The annotators were also instructed to correct punctuation marks. Using the alignment

	ZAEBUC-W1		ZAEBUC-W2	
	AR	**EN**	**AR**	**EN**
Keep	75%	82%	83%	88%
Replace	22%	11%	14%	6%
Insert	2%	5%	1%	4%
Delete	1%	2%	1%	1%

Table 5: The edit operations distributions in the ZAEBUC-Written Corpora.

algorithm developed by Alhafni et al. (2023), we derive and summarize the edit operations in Table 5.

We computed spelling and grammar correction IAA by comparing two independently corrected versions of a subset of texts: 11 for English and 9 for Arabic. Agreement was measured using normalized character-level Levenshtein similarity and word-level overlap. For Arabic, character-level similarity averaged 98.3% and word-level overlap 95%, while for English they were 98.7% and 97.2%, respectively. Most disagreements were non-erroneous corrections (80.9% for English and 61.7% for Arabic), including punctuation differences or valid but unnecessary corrections. The remaining disagreements involved errors such as incorrect casing, pronouns/articles/gender mismatches, or broken structure. Overall, these results give us confidence in the correction quality.

Morphological Annotation The corrected text versions of both ZAEBUC-Written corpora are manually annotated for tokenization, lemmatization, and Universal Dependency (UD) POS tagging (Nivre et al., 2017). For Arabic tokenization, we follow the Penn Arabic Tree Bank (Maamouri et al., 2004) scheme, which segments all clitics except the definite article, while for English tokenization, we only handle contractions. Lemmatization abstracts over inflectional variations of a lexical item with the same derivation and POS. For instance, the English verb forms go, goes, going, and gone are all lemmatized to go; similarly, the Arabic verb forms سيذهب *syðhb*[2] 'he will go', ذهبوا *ðhbwA* 'they went', and ذهبت *ðhbt* 'she went' are lemmatized to ذهب *ðhb*. We release Arabic lemmas in their *diacritized* form.

We computed IAA using 11 English texts (286 average words/text) and 9 Arabic texts (238 average words/text), comparing annotations from two annotators. In English, IAA was 99.9%, 99.6%, and 99.4% for tokenization, lemmatization, and POS tagging, respectively. In Arabic, the corresponding scores were 100%, 99.8%, and 99.4%, indicating very high agreement.

[2]Arabic transliteration in the HSB scheme (Habash et al., 2007).

3.2. ZAEBUC-Spoken

In Hamed et al. (2024), the ZAEBUC-Spoken corpus was collected as a multilingual multidialectal Arabic-English speech corpus. Hamed et al. (2024) make the recordings available with manual transcriptions and automatic morphological annotations. However, automatic annotations are insufficient for morphological analysis, as existing tools struggle with dialectal variations and code-switching, particularly in speech. To address this limitation, we extend their work by providing manual morphological annotations, including tokenization, lemmatization, and POS tagging, following the guidelines used in ZAEBUC-Written corpora.

We computed IAA by double annotating a subset of the data: 6,623 words in English and 2,343 words in Arabic. In English, the IAA was 99.6%, 99.4%, and 99.1% for tokenization, lemmatization, and POS tagging, respectively. In Arabic, the corresponding scores were 99.5%, 97.3%, and 99.6%. Overall, the results indicate very high agreement.

4. Benchmarks

Experimental Splits For ZAEBUC-W1, we use the splits from Alhafni et al. (2023), dividing the dataset into Train (70%), Dev (15%), and Test (15%), with balanced CEFR levels. We follow the same approach for ZAEBUC-W2. For ZAEBUC-Spoken, we use the splits from the original release Hamed et al. (2024), where Dev and Test sets were designed to minimize speaker overlap. Split statistics are in Table 14 (Appendix A).

4.1. Grammatical Error Correction

Models For Arabic GEC, we use the publicly available state-of-the-art sequence-to-sequence (Seq2Seq) model of Alhafni et al. (2023), which integrates morphological features and grammatical error detection information. It was trained on three Arabic GEC datasets: QALB-2014 (Mohit et al., 2014), QALB-2015 (Rozovskaya et al., 2015), and ZAEBUC-W1 (Habash and Palfreyman, 2022). For English GEC, we use the publicly available text-editing GECToR (Omelianchuk et al., 2020) model.

For both Arabic and English, we also benchmark four large language models (LLMs): two commercial models, GPT-4o and GPT-5 (OpenAI et al., 2024), and two open-source, Arabic-centric models, Jais-30B-Chat (Sengupta et al., 2023) and Fanar (Team et al., 2025). Following Alhafni and Habash (2025), we adopt their best performing prompting setup and use few-shot prompting with four examples. We use English prompts for GPT-4o and GPT-5, and Arabic prompts for Fanar and Jais, designed to elicit minimal edit-style corrections (Table 17 in Appendix C).

	ZAEBUC-W1 (AR)						ZAEBUC-W2 (AR)					
	Dev			Test			Dev			Test		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Seq2Seq	**87.6**	73.9	**84.5**	85.2	73.7	82.6	81.1	71.6	79.0	80.0	58.1	74.4
GPT-5	81.7	**77.2**	80.8	81.5	**76.2**	80.4	76.3	**77.9**	76.6	73.9	63.1	71.4
GPT-4o	86.5	76.1	84.2	**86.8**	76.1	**84.4**	**87.5**	75.9	**84.9**	**85.0**	**68.5**	**81.1**
Fanar	50.4	10.2	28.3	57.9	16.3	38.4	47.2	25.8	40.5	35.0	3.3	12.1
Jais	44.2	13.8	30.7	48.1	10.4	27.9	38.0	12.5	27.0	36.1	4.9	16.0

Table 6: GEC results on the Arabic ZAEBUC-Written Corpora in terms of precision, recall, and $F_{0.5}$.

	ZAEBUC-W1 (EN)						ZAEBUC-W2 (EN)					
	Dev			Test			Dev			Test		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
GECToR	57.1	37.5	51.7	55.0	37.6	50.3	51.8	36.5	47.8	57.6	38.9	52.5
GPT-5	**71.6**	**70.8**	**71.4**	**69.1**	**68.2**	**68.9**	56.9	**61.5**	57.8	62.6	**64.7**	63.0
GPT-4o	69.9	65.7	69.0	67.7	63.4	66.8	**60.9**	59.8	**60.7**	**67.4**	60.4	**65.9**
Fanar	15.1	23.1	16.2	11.7	20.0	12.8	8.0	17.9	9.0	11.0	26.5	12.5
Jais	27.4	36.4	28.8	28.0	34.6	29.2	16.4	29.6	18.0	21.5	29.3	22.7

Table 7: GEC results on the English ZAEBUC-Written Corpora in terms of precision, recall, and $F_{0.5}$.

Evaluation For Arabic GEC evaluation, we use the MaxMatch (M^2) scorer (Dahlmeier and Ng, 2012), while for English GEC evaluation, we use ERRANT (Bryant et al., 2017). Both the M^2 scorer and ERRANT compare the edits made by the GEC system to the annotated reference edits, calculating precision (P), recall (R), and $F_{0.5}$ scores. $F_{0.5}$ is commonly used in GEC and it weighs precision twice as much as recall, emphasizing the accuracy of system edits. To obtain the gold standard edits for Arabic, we apply the alignment algorithm proposed by Alhafni et al. (2023), whereas for English, we rely on the alignment algorithm built into ERRANT. Since GECToR does not support document-level GEC, we segment the English essays into sentences by splitting on periods, question marks, and exclamation marks. During the evaluation, we put the corrected segments back together to run the evaluation at the essay level.

Results We present the GEC benchmarking results in Tables 6 and 7. For Arabic GEC, the Seq2Seq model of Alhafni et al. (2023) achieves the best results on ZAEBUC-W1 Dev, which is expected since it was trained on ZAEBUC-W1 data. For the ZAEBUC-W1 Test, and Dev and Test sets of ZAEBUC-W2, GPT-4o yields the best performance.

For English GEC, all models generally perform better on ZAEBUC-W1 than on ZAEBUC-W2, with GPT-5 achieving the highest performance on ZAEBUC-W1 and GPT-4o leading on ZAEBUC-W2. Overall, the Arabic GEC results are better than the English results across both corpora, likely because the writers are native Arabic speakers.

4.2. Automated Essay Scoring

Models We treat the task of automated essay scoring (AES) as a text classification problem. For Arabic AES, we fine-tune CAMeLBERT MSA (Inoue et al., 2021a) on the training splits of both ZAEBUC-W1 and ZAEBUC-W2. Similarly, for English AES, we fine-tune BERT (Devlin et al., 2019) on the training splits of both ZAEBUC-W1 and ZAEBUC-W2. For both Arabic and English AES, the models were trained by using the average CEFR gold labels. Essays labeled as Unassessable are excluded from training but penalized during evaluation. Similar to GEC, we benchmark GPT-4o, GPT-5, Fanar, and Jais-30B-Chat using a comparable setup, employing few-shot prompts with four examples in English for GPT models and in Arabic for Fanar and Jais (Table 18 in Appendix C).

Evaluation We use Quadratic Weighted Kappa (QWK) (Cohen, 1968) as our primary evaluation metric, as it is standard in AES research (Ke and Ng, 2019). We also report macro precision (P), recall (R), and F_1 scores. Model predictions are evaluated against the averaged gold CEFR labels and in a multi-reference setting against the three individual gold labels per essay.

Results We present the AES results against the average gold CEFR labels in Tables 8 and 9. For Arabic, performance on ZAEBUC-W2 generally exceeds that on ZAEBUC-W1 across all metrics, with CAMeLBERT MSA achieving the highest QWK on ZAEBUC-W1, and GPT-5 obtaining the best QWK

	ZAEBUC-W1 (AR)								ZAEBUC-W2 (AR)							
	Dev				Test				Dev				Test			
	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1
BERT	**33.2**	26.8	30.8	28.3	**41.1**	31.3	35.0	32.6	47.4	**52.8**	50.0	46.8	54.5	**54.5**	**52.4**	**50.2**
GPT-5	25.2	23.9	37.9	23.6	25.0	**44.4**	37.5	**39.5**	**68.3**	51.2	**58.5**	**52.4**	**72.7**	33.3	28.6	29.5
GPT-4o	-3.4	30.6	29.5	27.1	-1.9	40.4	**48.6**	37.2	40.5	50.0	51.8	30.3	44.9	36.4	21.4	20.1
Fanar	25.0	**53.0**	**51.9**	**37.1**	3.2	24.1	25.9	14.0	47.5	29.1	16.8	15.6	39.5	8.3	11.9	9.8
Jais	11.7	15.5	19.0	6.9	-20.5	13.2	19.8	6.4	7.1	27.3	28.1	9.7	16.3	8.3	2.4	3.7

Table 8: AES results on the Arabic ZAEBUC-Written Corpora in terms of Quadratic Weight Kappa (QWK), and macro precision, recall, and F_1.

	ZAEBUC-W1 (EN)								ZAEBUC-W2 (EN)							
	Dev				Test				Dev				Test			
	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1
BERT	**62.1**	38.2	40.7	39.3	**81.9**	**48.6**	**48.9**	**48.6**	57.7	30.1	30.7	30.4	**33.7**	40.6	52.1	44.7
GPT-5	44.3	**54.5**	41.5	**45.0**	72.7	45.9	42.1	42.9	**60.4**	**53.8**	**52.1**	**52.7**	24.5	**51.9**	50.0	**47.8**
GPT-4o	46.5	51.0	**42.9**	**45.0**	70.6	40.1	40.5	39.6	45.3	34.7	38.7	34.6	28.2	43.2	**56.4**	44.6
Fanar	42.4	18.8	22.3	16.1	57.6	32.0	31.4	21.5	24.5	6.2	15.5	8.7	14.4	26.8	28.0	15.9
Jais	-1.9	0.7	14.3	1.3	12.0	9.8	22.6	8.3	1.9	17.2	17.3	2.3	-7.2	0.6	8.3	1.1

Table 9: AES results on the English ZAEBUC-Written Corpora in terms of Quadratic Weight Kappa (QWK), and macro precision, recall, and F_1.

on ZAEBUC-W2. In contrast, for English, models perform better on ZAEBUC-W1 than on ZAEBUC-W2. BERT achieves the best QWK on ZAEBUC-W1 and on ZAEBUC-W2 Test, while GPT-5 leads only on ZAEBUC-W2 Dev.

Notably, across both the Arabic and English written corpora, QWK and macro F_1 exhibit weak correlation, indicating that models producing consistent ordinal predictions are not necessarily those achieving the most accurate class-level classifications.

Multi-reference AES results are presented in Appendix B.

4.3. Morphological Tagging

Models For morphological tagging in the ZAEBUC-Written corpus, we apply Stanza (Qi et al., 2020) for English, and the MSA BERT-based morphological disambiguation models available in CAMeL Tools for Arabic (Obeid et al., 2020; Inoue et al., 2021b). In the case of the ZAEBUC-Spoken corpus, the automatic annotations are provided by Hamed et al. (2024) as part of their release, where they used Stanza for English words, and the MSA and Gulf Arabic morphological disambiguation models from CAMeL Tools for Arabic words. All of these models are used to provide tokenization, lemmas (diacritized for Arabic), and UD POS tags.

We also benchmark GPT-4o and GPT-5. For written data, we use few-shot prompting with four examples similar to the GEC and AES setups, while for spoken data, we use zero-shot prompting due to the lack of training splits (Table 19 in Appendix C). We do not report results for Fanar and Jais-30B-Chat, as their outputs were extremely noisy and did not yield meaningful results. To the best of our knowledge, these are the first reported results on Arabic morphological tagging using LLMs.

Among the previous annotation decisions for ZAEBUC-W1 and ZAEBUC-Spoken, we revise the annotation of *demonstrative pronouns*, which were previously labeled as 'DET' following the UD mapping provided by Taji et al. (2017). Instead, we annotate these cases as 'PRON' and update both the manual and automatic annotations accordingly. We also update the releases of ZAEBUC-W1 and ZAEBUC-Spoken to ensure consistency within ZAEBUC*.

Results We report results in terms of accuracy ZAEBUC-Written corpora in Tables 10 and 11, and ZAEBUC-Spoken corpus in Table 12.

On the written corpora, *tokenization* (Tok) results are comparable for Arabic and English, ranging from 98.1 to 99.8, with GPT-5 achieving the best performance in nearly all cases, except for the English test set of ZAEBUC-W1 where GPT-4o slightly outperforms it. For *POS tagging*, both languages show similar performance, with accuracies ranging from 94.8 to 97.3, again with GPT-5 leading overall. For *lemmatization* (Lex), GPT-5 consistently attains the highest accuracy across the majority of the datasets, reaching up to 99.1 for English and 95.4 for Arabic. English lemmatization results

	ZAEBUC-W1 (AR)						ZAEBUC-W2 (AR)					
	Dev			Test			Dev			Test		
	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex
CAMeL	99.2	95.8	94.3	99.3	95.1	94.4	99.2	95.5	92.5	99.2	95.1	94.1
GPT-5	**99.6**	**96.9**	**95.4**	**99.4**	**96.5**	**94.5**	**99.5**	**97.3**	94.9	**99.7**	**96.7**	**94.7**
GPT-4o	98.4	95.1	94.9	98.1	94.8	94.3	98.8	96.5	**95.1**	98.7	95.4	94.4

Table 10: Morphological tagging results on the Arabic ZAEBUC-Written Corpora in terms of accuracy across tokenization (Tok), part-of-speech tagging (POS), and lemmaization (Lex).

	ZAEBUC-W1 (EN)						ZAEBUC-W2 (EN)					
	Dev			Test			Dev			Test		
	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex
Stanza	99.1	95.0	96.7	99.1	94.8	96.5	99.5	95.7	97.5	99.5	95.7	97.4
GPT-5	**99.8**	**96.2**	**99.1**	99.7	**96.3**	**99.1**	**99.8**	**96.3**	**98.7**	**99.8**	**96.1**	**98.9**
GPT-4o	99.7	95.7	99.0	**99.8**	95.9	99.0	99.7	95.3	98.4	99.7	95.1	98.8

Table 11: Morphological tagging results on the English ZAEBUC-Written Corpora in terms of accuracy across tokenization (Tok), part-of-speech tagging (POS), and lemmaization (Lex).

	ZAEBUC-S (AR)						ZAEBUC-S (EN)						ZAEBUC-S (CSW)					
	Dev			Test			Dev			Test			Dev			Test		
	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex	Tok	POS	Lex
C&S	**89.0**	**87.7**	**62.5**	**89.9**	**86.3**	**64.8**	**96.9**	**95.8**	**98.7**	**97.3**	**96.2**	**98.5**	**89.1**	**87.3**	**69.2**	**90.2**	**87.0**	**72.1**
GPT-5	85.4	73.2	28.7	85.9	75.2	30.1	96.7	90.5	94.6	97.3	91.8	94.9	89.4	73.9	51.5	90.0	77.7	55.3
GPT-4o	80.1	72.7	28.0	81.3	72.8	29.4	89.4	88.5	93.1	91.0	90.5	94.3	84.8	74.5	49.4	85.4	77.3	53.0

Table 12: Morphological tagging results on the ZAEBUC-Spoken (ZAEBUC-S) Corpora in terms of accuracy across tokenization (Tok), part-of-speech tagging (POS), and lemmaization (Lex). C&S denotes CAMeL Tools and Stanza.

are generally higher than Arabic, with an absolute difference ranging from 2.1 to 5 points across splits.

On ZAEBUC-Spoken, performance drops across all morphological tagging tasks compared to the written corpora, most notably for Arabic and code-switched (CSW) data. This deterioration can be attributed to disfluencies and repetitions that disrupt the word flow. For Arabic, the task is further complicated as the spoken corpus includes both dialectal Arabic and MSA, whereas the written corpus is entirely MSA. CAMeL Tools and Stanza (C&S) outperforms GPT-5 and GPT-4o across all tagging tasks. Notably, the Arabic and code-switched lemmatization accuracy for the LLMs is significantly lower compared to C&S because they fail to produce diacritized Arabic lemmas.

4.4. Automatic Speech Recognition

Models We benchmark the ZAEBUC-Spoken corpus using Whisper (large-v3, 1.54B) (Radford et al., 2023). The spoken corpus provides two types of recordings for meetings; one audio file having all participants' audio streams (combined file) as well as separate files for each participant (separate files). However, it is worth noting that all publicly available recorded meetings have combined recording files but some are missing the separate recording files. We report evaluation results on the combined and separate recordings.

Evaluation We report ASR performance using Word Error Rate (WER). We evaluate Whisper's performance with and without being provided with the language flag when generating transcriptions. First, we obtain transcriptions without specifying the language of the utterances, to assess Whisper's performance in real-life scenarios, where the language may not be known beforehand. For monolingual Arabic and English utterances where Whisper fails to detect the correct language, we regenerate the transcriptions with providing the correct language tags. This allows us to evaluate Whisper's speech recognition performance independently of its language identification capabilities. We also evaluate Whisper's performance on language identification for Arabic and English monolingual utterances by providing F_1 scores.

	Separate Audio Recordings								Combined Audio Recordings							
	Without language flag				With language flag				Without language flag				With language flag			
	AR	EN	CSW	All	AR	EN	CSW	All	AR	EN	CSW	All	AR	EN	CSW	All
Dev	41.4	12.9	49.6	27.2	37.2	12.8	49.6	26.3	41.0	17.1	51.4	31.3	39.0	17.0	51.4	30.8
Test	39.4	13.7	48.0	29.9	36.2	13.7	48.0	28.9	40.6	15.5	49.1	29.3	37.4	15.0	49.1	28.2

Table 13: WER results of the ZAEBUC-Spoken corpus. We present results using separate and combined recordings for both setups, with and without the language flag. WER is reported across all utterances, as well as separately for Arabic (AR), English (EN), and Arabic-English code-switched (CSW) utterances.

Results With regards to Whisper's performance, we observe substantial drops in Arabic compared to English across both language identification and speech recognition tasks. For language identification, Whisper achieves F_1 scores of 75.2/73.4 and 90.7/89.6 on Arabic and English monolingual utterances, respectively, for Dev/Test sets. We present ASR results in Table 13. Even with providing the language flag, WER on Arabic remains high at around 36.2-39.0, as opposed to a significantly lower range of 12.8-17.0 for English. We note that due to the difference in number of utterances between the combined and separate recordings, we do not compare these two settings. In general, the results highlight the necessity of advancing ASR in Arabic and code-switching settings.

5. Conclusion and Future Work

We presented ZAEBUC*, a bilingual bimodal dataset that builds upon and enriches prior corpora with new annotations. The corpus covers Arabic and English in both written and spoken forms, where all the data is collected from a similar population. Our corpus supports a range of NLP tasks including grammatical error correction (GEC), automated essay scoring (AES), morphological tagging, and automatic speech recognition (ASR). We demonstrated the usability of our corpus by presenting benchmarking results of various NLP models, including LLMs, across different tasks, languages, and modalities. We hope that this corpus serves as a valuable resource to encourage further research in Arabic and English NLP.

In future work, we plan to further enrich ZAEBUC* with additional annotations including more morphological features. Additionally, we plan to investigate the interplay between the various tasks supported by ZAEBUC*, exploring how performance in one task (e.g., GEC) may influence or improve performance in others (e.g., AES or morphological tagging). Furthermore, we intend to expand the dataset to include more diverse user populations, domains, and dialects, thereby enhancing its utility for both monolingual and bilingual NLP research.

Acknowledgments

ZAEBUC was funded by Zayed University Research Incentive Fund awards (#R19068 and #R21072). We thank Dr. Michael Bowles for his project management help and Ramy Eskander for helpful discussions and his work at Ramitechs.

Limitations

ZAEBUC* offers a rich and diverse dataset, but a notable limitation is its relatively small size compared to the larger corpora commonly used in NLP research. This may present challenges for training models that require vast amounts of data. However, this limitation is balanced by the dataset's high-quality and diverse annotations, which provide a strong foundation for evaluating related tasks and studying their interactions. The bilingual and bimodal nature of the dataset also offers valuable insights that are not easily found in larger but less specialized corpora. Moreover, our benchmark experiments rely on closed-source commercial LLMs, which are subject to periodic updates that are not publicly documented. This introduces some uncertainty and may affect the reproducibility of our results over time.

Ethics Statement

In conducting this research and building ZAEBUC*, we adhered to strict ethical guidelines to ensure the protection, privacy, and well-being of all participants involved. Data collection was approved by the Zayed University's Research Ethics Committee, and all participants provided informed consent, ensuring their awareness of the research purpose, data usage, and the right to withdraw at any time without consequence. We ensured that all collected data from students was anonymized to protect their privacy and confidentiality. All annotators involved in the project were compensated fairly. We ensured that their efforts were recognized, and that the compensation reflected the time, expertise, and complexity of the tasks they undertook.

6. Bibliographical References

Bashar Alhafni and Nizar Habash. 2025. Enhancing text editing for grammatical error correction: Arabic as a case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17892–17914, Vienna, Austria. Association for Computational Linguistics.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

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

Vivian Cook. 2016. *Premises of multi-competence*, Cambridge Handbooks in Language and Linguistics, page 1–25. Cambridge University Press.

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

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

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.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

Luis Fernando D'Haro and Rafael E. Banchs. 2016. Automatic correction of asr outputs by using machine translation. In *Interspeech 2016*, pages 3469–3473.

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, 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.

Injy Hamed, Fadhl Eryani, David Palfreyman, and Nizar Habash. 2024. ZAEBUC-spoken: A multilingual multidialectal Arabic-English speech corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17770–17782, Torino, Italia. ELRA and ICCL.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*, pages 119–130.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of COLING*, pages 4561–4585.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021a. 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.

Go Inoue, Salam Khalifa, and Nizar Habash. 2021b. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of ACL*, pages 1708—-1719.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. 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 2274–2285, Online. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phuong Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0.

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 pro-

cessing. In *Proceedings of LREC*, pages 7022–7032.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of ACL: System Demonstration*, pages 101–108.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025a. Enhancing Arabic automated essay scoring with synthetic data and error injection. In *Proceedings of*

the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 549–563, Vienna, Austria. Association for Computational Linguistics.

Chatrine Qwaider, Kirill Chirkunov, Bashar Alhafni, Nizar Habash, and Ted Briscoe. 2025b. Evaluating prompt relevance in Arabic automatic essay scoring: Insights from synthetic and real-world data. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 162–178, Suzhou, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal dependencies for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.

Tomohiro Tanaka, Ryo Masumura, Hirokazu Masataki, and Yushi Aono. 2018. Neural error corrective language models for automatic speech recognition. In *Interspeech 2018*, pages 401–405.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

7. Language Resource References

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, Bangkok, Thailand. Association for Computational Linguistics.

May Bashendy, Walid Massoud, Sohaila Eltanbouly, Salam Albatarni, Marwan Sayed, Abrar Abir, Houda Bouamor, and Tamer Elsayed. 2026. Laila: A large trait-based dataset for arabic automated essay scoring.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022. Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2672–2681, Marseille, France. European Language Resources Association.

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.

Injy Hamed, Fadhl Eryani, David Palfreyman, and Nizar Habash. 2024. ZAEBUC-spoken: A multilingual multidialectal Arabic-English speech corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17770–17782, Torino, Italia. ELRA and ICCL.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A. Data Statistics

	ZAEBUC-W1				ZAEBUC-W2				ZAEBUC-S					
	AR		EN		AR		EN		AR		EN		CSW	
	Docs	Toks	Docs	Toks	Docs	Toks	Docs	Toks	Utts	Toks	Utts	Toks	Utts	Toks
Train	150	24,159	272	67,907	73	15,460	216	48,922	-	-	-	-	-	-
Dev	33	5,100	58	15,206	16	2,999	46	10,780	1,186	14,971	1,611	31,585	629	14,386
Test	31	4,976	58	14,365	15	2,899	46	10,307	649	8,355	880	16,973	301	6,871

Table 14: Dataset statistics of the ZAEBUC-Written (W1 and W2) and ZAEBUC-Spoken (S) corpora in terms of number of documents (Docs), tokens (Toks), and utterances (Utts) across the Train, Dev, and Test splits. ZAEBUC-Spoken has only Dev and Test splits.

B. Automated Essay Scoring Multi-Reference Evaluation

Tables 15 and 16 present AES multi-reference evaluation results against the three individual gold CEFR labels per essay. Compared to evaluation against the average CEFR label §4.2, the model rankings shift toward LLMs. Under the average-label setting, BERT performs best on most evaluation sets, particularly on ZAEBUC-W1 for both Arabic and English (Tables 8 and 9). In contrast, under the multi-reference evaluation, LLMs achieve the best performance on most sets. This pattern likely reflects the ability of LLMs to capture multiple valid scoring patterns, whereas BERT aligns more closely with a single averaged label, as it was trained on the average label.

	ZAEBUC-W1 (AR)								ZAEBUC-W2 (AR)							
	Dev				Test				Dev				Test			
	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1
BERT	65.9	43.9	46.9	45.3	56.1	36.0	40.0	37.9	71.4	**91.7**	**83.3**	**85.5**	64.0	57.6	55.6	54.7
GPT-5	**88.1**	**88.8**	**92.1**	**89.4**	**87.2**	73.9	84.1	76.4	**81.0**	58.9	65.2	59.5	**93.3**	**97.2**	**95.0**	**95.8**
GPT-4o	13.7	45.4	44.4	44.7	84.7	**92.8**	**91.1**	**90.3**	50.0	63.3	59.7	47.1	49.0	68.7	68.3	55.4
Fanar	71.8	79.6	80.4	71.1	64.9	63.2	61.7	52.5	56.5	52.7	30.8	33.8	61.5	49.3	56.0	52.0
Jais	22.7	39.7	32.6	27.0	-12.1	23.9	20.6	16.5	7.1	27.3	28.1	9.7	25.7	31.8	43.3	28.3

Table 15: Multi-reference AES results on the Arabic ZAEBUC-Written Corpora in terms of Quadratic Weight Kappa (QWK), and macro precision, recall, and F_1.

	ZAEBUC-W1 (EN)								ZAEBUC-W2 (EN)							
	Dev				Test				Dev				Test			
	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1	QWK	P	R	F_1
BERT	82.1	71.3	74.2	72.7	**98.1**	**99.0**	**97.2**	**98.0**	87.0	62.2	65.0	63.5	45.4	46.6	57.9	50.6
GPT-5	94.5	**98.4**	93.3	95.2	94.9	95.0	95.0	95.0	**93.9**	**86.7**	**96.2**	**89.2**	**94.8**	**98.8**	**96.7**	**97.5**
GPT-4o	**95.5**	96.9	**96.8**	**96.7**	95.6	96.5	96.2	96.2	84.5	62.3	63.2	62.0	38.9	49.0	61.4	51.7
Fanar	80.6	75.0	54.8	56.7	85.7	68.0	50.7	50.3	34.9	20.2	24.2	22.0	17.3	35.9	30.2	24.8
Jais	-1.8	14.3	18.8	8.8	17.9	18.5	24.5	16.2	1.8	17.8	17.3	3.3	-7.6	1.2	11.1	2.2

Table 16: Multi-reference AES results on the English ZAEBUC-Written Corpora in terms of Quadratic Weight Kappa (QWK), and macro precision, recall, and F_1.

C. Prompts

Task	Prompt

GEC (EN)	You are an English grammatical error correction tool.
	Your task is to identify and correct **only grammatical and spelling errors** in English sentences, while preserving their original meaning and phrasing.
	You will receive examples and inputs in JSON format and must always return a JSON object with the schema:
	{
	"input": "<original sentence>",
	"output": "<corrected sentence>"
	}
	If either the input or output is missing, your answer is invalid.
	Guidelines:
	1. Make the minimal edits necessary to correct grammar or spelling.
	2. Do not rephrase correct parts of the sentence.
	3. Avoid altering the meaning by adding or removing information.
	4. Output **only** valid JSON, no extra text, comments, or explanations.
	Return a JSON object with both "input" and "output" fields.
	The "input" field must contain the original sentence,
	and the "output" field must contain the corrected sentence.
	Here is the sentence:

GEC (AR)	أنت أداة لتصحيح الأخطاء النحوية والإملائية في اللغة العربية.
	مهمتك هي تحديد وتصحيح **الأخطاء النحوية والإملائية فقط** في الجمل العربية مع الحفاظ على المعنى الأصلي للسياق دون أي تغيير في المقصود.
	سيتم تزويدك ببعض الأمثلة والمدخلات على شكل JSON، ويجب عليك إخراج النتائج في صيغة JSON بنفس البنية التالية:
	{
	"input": "<الجملة الأصلية>",
	"output": "<الجملة المصححة>"
	}
	إذا كان أي من الحقلين input أو output مفقودا، فإجابتك غير صالحة.
	التعليمات:
	1. أجر أقل عدد ممكن من التعديلات لتصحيح الجملة.
	2. لا تعد صياغة الأجزاء الصحيحة نحويا أو أسلوبيا.
	3. تجنب تغيير المعنى من خلال إضافة أو حذف أي المعلومات.
	4. أخرج النتيجة بصيغة JSON فقط، دون أي نص إضافي أو شرح أو علامات تنسيق.
	أرجع كائن JSON يحتوي على الحقلين "input" و "output".
	يجب أن يحتوي الحقل "input" على الجملة الأصلية، بينما يجب أن يحتوي الحقل "output" على الجملة المصححة.
	إليك الجملة:

Table 17: Prompts used for the GEC experiments, with English (EN) prompts for English data and Arabic (AR) prompts for Arabic data.

Task	Prompt

AES (EN)	You are an English essay scoring system.
	Your task is score each essay according to the CEFR (Common European Framework of Reference) writing scale.
	The CEFR scale uses six levels: A1, A2, B1, B2, C1, C2. Here are the guidelines for each level:
	- A1 (Beginner): Can write simple isolated phrases and sentences.
	- A2 (Elementary): Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.
	- B1 (Intermediate): Can write straightforward connected texts on a range of familiar subjects by linking a series of phrases.
	- B2 (Upper-Intermediate): Can write clear, detailed texts on a variety of subjects, synthesising and evaluating information and arguments from a number of sources.
	- C1 (Advanced): Can write clear, well-structured, and detailed text on complex subjects, expressing their views in a way that shows a mastery of text-forming strategies.
	- C2 (Proficient): Can write clear, well-structured, and detailed text on complex subjects, mastering a variety of text-forming strategies. At this expert level, a user can write with fluency and accuracy.
	You will receive examples and inputs in JSON format and must always return a JSON object with the schema:
	{
	"input": "<input sentence>",
	"output": "<CEFR label>"
	}
	If either the input or output is missing, your answer is invalid.
	Guidelines:
	1. Make sure to follow the CEFR guidelines above.
	2. Output **only** valid JSON, no extra text, comments, or explanations.
	3. The output label must be one of the six CEFR levels.
	Return a JSON object with both "input" and "output" fields.
	The "input" field must contain the input sentence, and the "output" field must contain the CEFR label.
	Here is the sentence:

AES (AR)	أنت نظام لتقييم المقالات المكتوبة باللغة العربية.
	مهمتك هي تقييم كل مقالة وفقا لمقياس الكتابة الخاص بإطار CEFR (Common European Framework of Reference).
	يستخدم مقياس CEFR ستة مستويات: A1، A2، B1، B2، C1، C2.
	فيما يلي الإرشادات الخاصة بكل مستوى:
	- A1 (مبتدئ): يستطيع كتابة عبارات وجمل بسيطة ومعزولة.
	- A2 (أساسي): يستطيع كتابة سلسلة من العبارات والجمل البسيطة المترابطة بروابط بسيطة مثل "و" و "لكن" و "بسبب".
	- B1 (متوسط): يستطيع كتابة نصوص مترابطة وبسيطة حول مواضيع مألوفة، من خلال ربط سلسلة من العبارات معا.
	- B2 (فوق المتوسط): يستطيع كتابة نصوص واضحة ومفصلة حول مجموعة متنوعة من المواضيع، مع القدرة على تلخيص وتقييم المعلومات والحجج من مصادر متعددة.
	- C1 (متقدم): يستطيع كتابة نصوص واضحة ومنظمة ومفصلة حول مواضيع معقدة، والتعبير عن آرائه بطريقة تظهر إتقانه لاستراتيجيات بناء النصوص.
	- C2 (متمكن): يستطيع كتابة نصوص واضحة ومنظمة ومفصلة حول مواضيع معقدة، مع إتقان مجموعة متنوعة من استراتيجيات بناء النصوص. في هذا المستوى المتقدم جدا، يمكن للكاتب الكتابة بطلاقة ودقة.
	ستتلقى أمثلة ومدخلات بصيغة JSON، ويجب عليك دائما إخراج كائن JSON بالبنية التالية:
	{
	"input": "<نص الإدخال>",
	"output": "<تسمية مستوى CEFR>"
	}
	إذا كان أي من الحقلين "input" أو "output" مفقودا، فإن إجابتك غير صالحة.
	التعليمات:
	1. تأكد من اتباع إرشادات CEFR الموضحة أعلاه.
	2. أخرج النتيجة بصيغة JSON فقط، دون أي نص إضافي أو شرح.
	3. يجب أن تكون التسمية الناتجة واحدة من مستويات CEFR الستة.
	أرجع كائن JSON يحتوي على الحقلين "input" و "output".
	يجب أن يحتوي الحقل "input" على الجملة الأصلية،
	بينما يجب أن يحتوي الحقل "output" على تسمية ال CEFR.
	إليك الجملة:

Table 18: Prompts used for the AES experiments, with English (EN) prompts for English data and Arabic (AR) prompts for Arabic data.

Task	Prompt

Morphological Tagging (EN)	You are an English morphological analyzer. Your task is provide the following morphological features for each input word in the given text:
	- the tokenized word
	- the part-of-speech (POS) tag according to the universal dependencies (UD) framework
	- the lemma of the word
	You will receive examples and inputs in JSON format and must always return a JSON object with the schema:
	{
	"input": "<input sentence>",
	"output": "<annotated sentence>"
	}
	Each word in the annotated sentence has the following format: <w>input_word<w>tokenized_word<w>pos_tag<w>lemma<w>. Your output must follow this exact format
	If either the input or output is missing from the output JSON, your answer is invalid.
	Guidelines:
	1. Make sure to follow the guidelines above.
	2. Output **only** valid JSON, no extra text, comments, or explanations.
	3. Each word in the output annotated sentence must follow the format above.
	4. The output annotated sentence **must** have the same number words as the input sentence.
	Return a JSON object with both "input" and "output" fields.
	The "input" field must contain the input sentence,
	and the "output" field must contain the annotated sentence where each word is annotated with the morphological features specified above.
	Here is the sentence:

Morphological Tagging (AR)	أنت محلل صرفي للغة العربية.
	مهمتك هي استخراج الخصائص الصرفية التالية لكل كلمة في النص المعطى:
	- الكلمة بعد التجزئة
	- القسم النحوي للكلمة وفق إطار universal dependencies
	- المدخل المعجمي للكلمة
	ستتلقى أمثلة ومدخلات بصيغة JSON، ويجب عليك دائما إخراج كائن JSON بالبنية التالية:
	{
	"input": "<الجملة الأصلية>",
	"output": "<الجملة الموسمة>"
	}
	كل كلمة في الجملة الموسمة يجب ان تكون بالصيغة التالية:
	<w>الكلمة المدخلة<w>الكلمة بعد التجزئة<w>القسم النحوي<w>المدخل المعجمي<w>
	ويجب ان يتبع الإخراج هذا الشكل تماما.
	اذا كان أي من الحقلين "input" أو "output" مفقودا من كائن JSON الناتج، فإن الإجابة غير صالحة.
	التعليمات:
	1. تأكد من اتباع التعليمات المذكورة أعلاه بدقة.
	2. أخرج النتيجة بصيغة JSON فقط، دون أي نص إضافي أو تعليقات أو شروحات.
	3. كل كلمة في الجملة الموسمة يجب ان تتبع الصيغة المحددة أعلاه.
	4. يجب ان تحتوي الجملة الموسمة على نفس عدد الكلمات الموجود في الجملة الأصلية.
	أرجع كائن JSON يحتوي على الحقلين "input" و "output".
	يجب ان يحتوي الحقل "input" على الجملة الأصلية،
	ويجب ان يحتوي الحقل "output" على الجملة الموسمة التي تم فيها تمييز كل كلمة بالخصائص الصرفية المحددة أعلاه.
	إليك الجملة:

Table 19: Prompts used for morphological tagging experiments, with English (EN) prompts for English data and Arabic (AR) prompts for Arabic data.