



Gazelle: An Instruction Dataset for Arabic Writing Assistance

Samar M. Magdy^ξ Fakhraddin Alwajih^{λ*} Sang Yun Kwon^{λ*}

Reem Abdel-Salam^φ Muhammad Abdul-Mageed^{λ,ξ,γ}

^λThe University of British Columbia ^ξMBZUAI

^γInvertible AI ^φCairo University

samar.magdy@mbzuai.ac.ae, muhammad.mageed@ubc.ca

Abstract

Writing has long been considered a hallmark of human intelligence and remains a pinnacle task for artificial intelligence (AI) due to the intricate cognitive processes involved. Recently, rapid advancements in generative AI, particularly through the development of Large Language Models (LLMs), have significantly transformed the landscape of writing assistance. However, underrepresented languages like Arabic encounter significant challenges in the development of advanced AI writing tools, largely due to the limited availability of data. This scarcity constrains the training of effective models, impeding the creation of sophisticated writing assistance technologies. To address these issues, we present *Gazelle*, a comprehensive dataset for Arabic writing assistance. In addition, we offer an evaluation framework designed to enhance Arabic writing assistance tools. Our human evaluation of leading LLMs, including GPT-4, GPT-4o, Cohere Command R+, and Gemini 1.5 Pro, highlights their respective strengths and limitations in addressing the challenges of Arabic writing. Our findings underscore the need for continuous model training and dataset enrichment to manage the complexities of Arabic language processing, paving the way for more effective AI-powered Arabic writing tools.

1 Introduction

Writing is an essential skill for both individuals and professionals, requiring a long-term commitment to mastery that can span years of continuous learning and refinement (Flower and Hayes, 1980; Du et al., 2022). The distinct cognitive processes involved in writing underscore the significant importance and complexity of writing as a task. Consequently, in the fields of AI and natural language generation (NLG), writing is considered a hallmark of human intelligence, and researchers have long viewed it as a pinnacle task for

*These authors contributed equally.

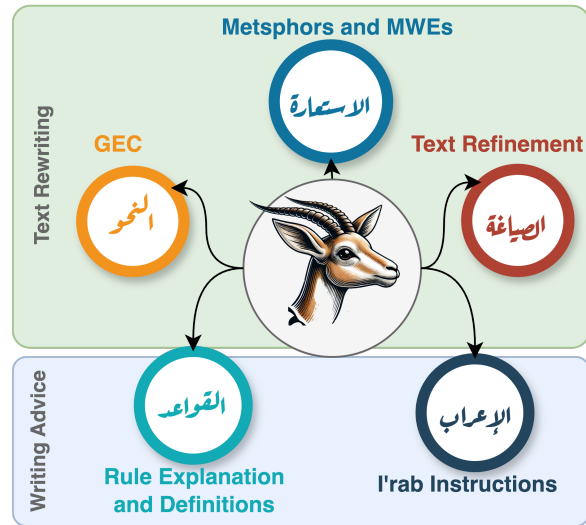


Figure 1: Task categorization for *Gazelle*. GEC: Grammatical Error Correction (التصحيح النحوي); MWEs: Multi-Word Expressions.

their studies (Ippolito et al., 2022; Urmeneta and Romero, 2024; Chen et al., 2024). In recent years, the landscape of writing assistance has been transformed by rapid advances in generative AI, particularly through the development of LLMs (Ippolito et al., 2022; Lee et al., 2024). These neural models, trained on the intricacies of language by processing vast amounts of text, have led to the creation of AI-powered tools designed to provide comprehensive writing assistance across a wide range of tasks (Zhang et al., 2022; Chowdhery et al., 2023). Models like GPT-3 and GPT-4 (Brown et al., 2020; Achiam et al., 2023) have demonstrated impressive capabilities in both understanding and generating text (Bubeck et al., 2023). These LLMs have significantly expanded the potential for "Human-AI co-writing," where AI systems provide flexible support that extends far beyond simple grammar and spell-checking to polishing texts and generating completely new content, significantly enhanc-

ing writing assistance for diverse writing tasks (Li et al., 2024).

However, underrepresented languages like Arabic often face substantial challenges in developing advanced AI tools, including writing assistance, largely due to the scarcity of relevant training data. In this work, we seek to bridge this gap by introducing *Gazelle*, a dataset specifically curated for Arabic writing, designed to advance the development of AI-powered writing tools for underrepresented languages. *Gazelle* is a manually curated instruction-style dataset crafted in both English and Arabic and organized around two primary themes. The first theme, **text rewriting**, aims at supporting learners in rewriting their texts across different linguistic levels. This encompasses (i) *Grammatical Error Correction* (GEC) at various levels such as orthographic, morphological, syntactic, and semantic with categories beyond those proposed in the Arabic Learner Corpus (ALC) (Alfaifi and Atwell, 2014) adopted in previous works. Text rewriting also involves corrections of deeper semantic levels beyond traditional GEC that cover (ii) *Metaphors and Multi-word Expressions* (MWEs) and (iii) *Text Refinement*, which focuses on improving texts by replacing dialectal sequences with their equivalent standard forms and fixing poor language.

The second theme, **writing advice**, covers (iv) *Rule Explanation and Definitions* as well as (v) *I'rab*. In Arabic, I'rab or declension refers to the variation in a word's form that indicates its grammatical case and role within a sentence. In *Gazelle*, the I'rab category guides learners through the different grammatical roles within a text, such as identifying when a word takes on a nominal case due to its role as a subject in a sentence. Across its various categories, *Gazelle* also provides explanations for the suggested corrections. Figure 1 outlines the main tasks covered in *Gazelle*, and Figure 2 illustrates each of these tasks.

Our contributions can be summarized as follows: (1) We introduce a comprehensive, manually curated dataset specifically for Arabic, covering two different writing themes across five distinct tasks. (2) We expand the Arabic Learner Corpus (ALC) (Alfaifi and Atwell, 2014) error taxonomy by developing a fine-grained classification, including adding sub-classes and detailed sub-subclasses under the main error categories. Furthermore, we generate synthetic data from the Arabic Tree Bank (ATB) (Maamouri et al., 2004) that aligns with our

new classification and sub-classes for specific error types in the ALC error taxonomy. (3) We conduct a comparative evaluation of current leading LLMs—GPT-4 (Brown et al., 2020), GPT-4o (Achiam et al., 2023), Cohere Command R+,¹ and Gemini 1.5 Pro (Reid et al., 2024)—focusing on their capabilities in different Arabic writing tasks.

The rest of this paper is organized as follows: In Section 2, we provide an overview of related work. Section 3 introduces our manually curated datasets for Arabic writing assistance. In Section 4, we provide a comparison of *Gazelle* with other GEC datasets. Section 5 presents our evaluation of the frontier LLMs on Arabic writing tasks. We also detail our human evaluation methodology and results. We conclude in Section 6.

2 Related Work

LLMs for Writing Assistance. The advent of LLMs and their enhancement through data augmentation and instruction tuning have significantly advanced automated writing assistance tools (Wei et al., 2022; Ouyang et al., 2022; Raheja et al., 2023). Models trained on extensive datasets of human-written instructions, using methods like instruction fine-tuning, have proven to enhance their ability to generalize across various tasks by leveraging large, diverse sets of instructional data (Zhang et al., 2023). Models like FLAN-T5 (Chung et al., 2024) and GPT-4 (Bubeck et al., 2023; Peng et al., 2023; Chung et al., 2024) demonstrate the progress and capabilities of incorporating a wide array of tasks during fine-tuning to achieve robust performance and adaptability in various writing applications (Sanh et al., 2021; Liu et al., 2022). We provide detailed examples of recent advancements in instruction-tuned models and their application on writing assistance in Appendix A.

Progress in Arabic Writing Tasks. Arabic writing tasks are complex due to its linguistic and morphological diversity, compounded by orthographic ambiguities from optional diacritics and dialectal variations (Abdul-Mageed et al., 2020; Belkebir and Habash, 2021a). Despite these challenges, significant progress has been made in terms of model development and dataset developments across different writing tasks such as GEC (Zaghouani et al., 2014; Habash and Palfreyman, 2022; Kwon et al.,

¹<https://docs.cohere.com/docs/command-r-plus>

Task	Error Class	Incorrect Sentence	Corrected Sentence	Explanation
GEC	Morphology	سجلت وحدات الحماية المدنية 999. حادث مرور أدت إلى وفاة 32 شخصا.	سجلت وحدات الحماية المدنية 999. حادث مرور أدى إلى وفاة 32 شخصا.	الخطأ في الفعل أدى، يعبر عن حدث أو موقف مذكر لذلك يجب أن يكون خالي من تاء التأنيث.
Metaphors	Metaphors	كان نقده كالمخ في العسل.	كان نقده كالعسل في العسل.	الخطأ في استخدام التعبير كالمخ في العسل، مما يغير المعنى المقصود من شيء ضار متنكر كشيء حلو إلى شيء غير متناسب ولكنه أقل ضرراً.
Multi Word Expressions	MWEs	وقد لمح بذلك إلى قيادة "الموساد المخابرات الإسرائيلية التي تريد أن تأخذ علي رأسها مسؤولية الحملة الدولية بإيران باعتباره موضوعاً أمنياً.	وقد لمح بذلك إلى قيادة الموساد المخابرات الإسرائيلية التي تريد أن تأخذ علي عاتقها مسؤولية الحملة الدولية ضد إيران باعتباره موضوعاً أمنياً.	الخطأ في كلمة بطيء: كتبت الهجزة المتطرفة على الياء لأن حركة الحرف الذي قبلها الكسرة.
Text Refinement	Style Transfer	وأنت تعرف جيداً، ولا عارف أن أصابعنا تحت ضرسه.	وأنت تعرف جيداً، وتعرف أن حياتنا بين يديه.	(أصابعنا تحت ضرسه) مصطلح مصري عامي بمعنى أنهم تحت رحمته أو تحكمه، ولا يصبح كتابته في الفصحى.
I'rab	Declension	اعرب الجملة التالية إعراباً كاملاً: الجملة: راقب الوالد أولاده وهم يلعبون.	بكل تأكيد، إليك إعراب الجملة بالكامل: راقب: فعل ماضٍ مبني على الفتح. الوالد: فاعل مرفوع وعلامة رفعه الضمة مبني في محل جر مضاف إليه. الواو: واو الحال وهم: ضمير رفع منفصل مبني في محل رفع مبتدأ. يلعبون: فعل مضارع مرفوع وعلامة رفعه ثبوت نون الإعراب في آخره خبر المبتدأ وهم: والجملة الاسمية من (هم يلعبون) في محل نصب حال.	

Figure 2: Examples of corrections and explanations for various Arabic writing tasks included in *Gazelle*. For instructions in English, see Table D.1

2023), text simplification (Hazim et al., 2022; Al-hafni et al., 2024) and text summarization (Al-Maleh and Desouki, 2020; Lagrini and Redjimi, 2021). Further details about the progress in Arabic writing tasks are provided in Appendix A.

Explainable Writing Systems. Recent advancements in explainable writing systems have greatly improved the transparency and alignment of tools with instructional goals across various tasks. In GEC, tools like ERRANT (Bryant et al., 2017a) provide clear error annotations and edit boundaries, enhancing users’ understanding of grammatical mistakes. In text simplification and feedback generation, systems strive to offer precise and relevant explanations, ensuring that users receive actionable and comprehensible feedback (Agarwal, 2022). These efforts highlight the importance of creating tools that deliver understandable insights and ensure alignment with instructional goals (Kunz and Kuhlmann, 2024). We provide further details on the progress of explainable writing systems in Appendix A. However, to the best of our knowledge, no prior work has investigated instruction-tuned LLMs’ ability to provide explanations and definitions for diverse writing tasks, organized around distinct pedagogical themes.

3 Gazelle

As mentioned in Section 1, *Gazelle* comprises two main themes, text rewriting and writing advice, across five distinct types of writing assistance instructions that we manually curate. *Gazelle* encompasses parallel data with instructions crafted in both Arabic and English, covering both input and

output, allowing users to fine-tune models for these various tasks bilingually. Except for the I’rab and Arabic grammatical rules and definitions datasets, we manually translate the Arabic explanations and instructions into English. To ensure thorough coverage, we curate instructions for each of these tasks separately from several publicly available online sources.² Table 1 presents the statistics for text rewriting tasks, while Table 2 provides statistics for tasks related to writing advice. Examples of these tasks are illustrated in Figure 2. For the English translations of the GEC explanations and definitions task dataset in Table 1, we leverage GPT-4o³ to accurately translate the manually curated Arabic content. This process involved iteratively prompting GPT-4o and refining the generated through multiple rounds of manual review to ensure the translations were both accurate and contextually appropriate, allowing us to produce high-quality English translations. We explain each of our instruction categories next.

3.1 Text Rewriting

3.1.1 GEC

Fine-Grained Error Taxonomy. The original ALC error taxonomy comprises 29 error tags categorized under five main error classes: *Orthographic*, *Morphology*, *Syntax*, and *Semantics*. Additionally, Belkebir and Habash (2021b) introduced two more error classes: *Split* and *Merge*. Building upon this foundation, we expand the orig-

²See Appendix G.2 for example sources we use.

³<https://openai.com/index/hello-gpt-4o/>

Task	Category	Source	Language	Total
GEC	Orthographic	Online Sources	Arabic	637
	Syntax	Online Sources		377
	Semantic	Online Sources		362
	Morphology	Online Sources		55
	Punctuation	Online Sources		28
	Split	ATB		100
	Merge	ATB		100
GEC_ATB		ATB	Arabic	1,433
GEC + Definitions		Online Sources	Arabic	290
GEC + Explanations		Online Sources	Arabic	1,659
MWE		PARSEME dataset	Arabic English	98
Text Refinement		In-house Project	Arabic English	770

Table 1: Data statistics for *text rewriting* tasks.

Task	Source	Language	Total
Rule Explanations & Definitions	Online Sources	Arabic	1,445
	Online Sources	English	1,005
Prab	Online Sources	Arabic	1,090

Table 2: Data statistics for *writing advice* tasks.

inal ALC error taxonomy to develop a fine-grained grammatical error taxonomy.

Our proposed taxonomy introduces two more sub-classes to the original ALC error categories: *Closed Class Error* under *Syntax* and *Special Expression Error* under *Semantic*, bringing the total to 31 error tags. *Closed class* errors address issues with closed class items, such as *the five nouns*, *the five verbs*, and *pronouns*. *Special expression errors* covers the use of collocations, multi-word expressions, and other commonly misused terms. Additionally, each of our new error categories include several sub-subclasses. For example, the sub-class *Number* under *Syntax* which pertains to the correct use of singular, dual, and plural forms in nouns, verbs, and pronouns is further divided into several sub-subclasses. Our newly-developed categorization scheme incorporates extensive knowledge of Arabic grammar, aiming to provide users with not only the category of grammatical error but also its sub- and sub-subclasses. Table D presents a detailed overview of all the newly introduced categories. It comprises 7 classes, each with an average of 9 sub-classes.⁴

GEC Instructions. We develop a set of 3,382 bilingual instructions for GEC, based on our newly extended error taxonomy. This comprehensive taxonomy includes a wide range of error sub-classes and sub-subclasses, aiding identification of specific

⁴We exclude all error tags within the *Other* error class (e.g., OO, XO, SO, MO) to avoid complications.

errors. We manually curate these sentences from various Arabic learning online sources to reflect the detailed error categories in our extended taxonomy. Each sentence pair includes grammatical errors along with their corrected versions and explanations, as shown in Table D.1. Examples of these detailed error sub-classes and their coverage are provided in Appendix C.

GEC Error Explanations. We introduce a dataset containing detailed explanations for each type of error sub-class and sub-subclass, comprising a total of 1,659 entries sourced from websites and books. This dataset enables users not only identify the type of error but also understand the explanations behind it. The explanations are provided in both Arabic and English. We provide examples of these detailed error explanations in Appendix C.

ATB Synthetic Data. ATB is a collection of texts annotated for morphosyntax with English glosses. We utilize all parts of the ATB to generate 100 examples with a total of 1,633 covering the *Orthographic*, *Syntax*, *Split*, *Merge*, *Morphological*, and *Punctuation* error classes along with their sub-classes. More details on the examples generated are provided in Appendix C.

Generating synthetic data from ATB allows us to efficiently acquire real-world data with human labels. We then use templates to format these examples into an instruction format.

For each error type, we follow specific rules that dictate how errors are created, ensuring consistency throughout the process. We prioritize aligning our error generation with those made by humans, rather than relying on random or arbitrary variations. For instance, with OM (Orthographic Missing) errors, where characters are missing, we simulate typical human mistakes by deliberately omitting characters that are commonly omitted in real-world usage. Appendix E provides detailed information on the methodology we follow to generate these synthetic errors from ATB.

3.1.2 Metaphors and MWEs

MWEs are diverse and arbitrary word combinations that co-occur frequently. They include idiomatic phrases, compound words, and other constructions with varying levels of transparency and fixedness, all sharing the characteristic of crossing word boundaries (Sag et al., 2002). Our dataset comprises 330 sentences featuring incorrect usage

of metaphors or MWEs, along with corrected versions and explanations of the errors. The sentences with metaphors were manually collected from online resources. For MWE sentences, we randomly selected 98 sentences from the PARSEME dataset (Mohamed et al., 2022) a collection of verbal multiword expressions in Arabic. To generate sentences with errors, we use GPT-4o to modify correct metaphors and MWEs into incorrect ones. We provide the correct sentence and instruct it to alter the metaphor or MWEs.

3.1.3 Text Refinement

Text Refinement refers to improving the quality of MSA by addressing informal or colloquial words in a text or by enhancing texts written in poor language. We create a set of instructions comprising a total of 770 sentences (we collect these sentences from an in-house project) focused on detecting informal words and improving text quality, along with explanations of these errors. Examples regarding this task can be found in Appendix C.

3.2 Writing Advice

3.2.1 Rule Explanation and Definitions

We develop a bilingual dataset consisting of 2,450 examples that provide explanations and definitions of grammatical rules in both Arabic and English. This dataset covers a wide range of topics in Arabic grammar, from basic concepts such as nouns, verbs, negation, and sentence structures to advanced topics like verb inflections, formations, and stylistic nuances. Each entry offers comprehensive definitions and illustrative examples to enhance understanding. Contextual examples are also included to demonstrate the practical application of these rules. Illustrative examples can be found in Figure D.1.

3.2.2 I'rab

I'rab, or declension, is a term in Arabic that refers to the system of nominal, adjectival, or verbal suffixes used to mark the grammatical case of a word. These suffixes help identify the function of a word within a sentence (Hamani, 2014). We design 1,090 instructions that provide the full I'rab of sentences, covering a wide range of topics in Arabic grammar, as detailed in Figure D.1.

4 Gazelle in Comparison

The QALB dataset comprises manually corrected Arabic texts, including MSA commentaries from Aljazeera by native speakers (L1) and texts by

non-native learners (L2) (Rozovskaya et al., 2014; Mostefa et al., 2015). The ZAEBUC corpus, on the other hand, contains essays written by first-year students at Zayed University in the UAE (Habash and Palfreyman, 2022). While both the QALB and ZAEBUC corpora include multiple errors per sentence or paragraph, facilitating automatic evaluation, our datasets are curated with a different focus.

We manually curate the *Gazelle* dataset to target specific error types within the ALC error taxonomy by extending the existing error categories and incorporating our own extended categories. Each sentence, sourced from various Arabic learning online sources, represents a specific subclass or sub-subclass of our extended error taxonomy and contains only one type of error along with explanation of this type of error. This targeted approach makes our dataset particularly useful for learners aiming to understand and correct specific types of errors. We provide a comparison of *Gazelle* dataset to the QALB and ZAEBUC corpora in Appendix F.1

5 Evaluating Leading LLMs on Gazelle

We evaluate *Gazelle* on current leading LLMs, namely GPT-4, GPT-4o, Gemini 1.5 Pro, and Cohere Command R+, to assess their capabilities across our proposed tasks in Arabic. We test the models under a zero-shot setting (Sanh et al., 2021), allowing us to assess the models' inherent capabilities in handling our proposed writing tasks. For each task, we initially tested the prompts and iteratively refined a zero-shot prompt tailored to each subtasks. Further details on our prompt design process, along with examples used for evaluation, are provided in Appendix F.

5.1 Human Evaluation

We conduct human evaluation to assess the performance of the aforementioned LLMs on *Gazelle*'s five subtasks. Developing evaluation criteria for these subtasks in Arabic writing was challenging due to the lack of existing standard measures. Consequently, we create a list of suitable measures for all subtasks and iteratively refine them. Each criterion was clearly defined to ensure both comprehensiveness and clarity.

For evaluation, we ask four native Arabic speakers to assess the models' performance by randomly sampling a subset of 20 sentences from each task,

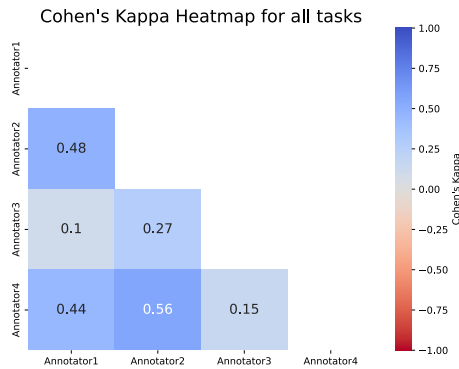


Figure 3: Overall inter-annotator agreement for human evaluation measured by Cohen’s Kappa.

resulting in a total of 100 sentences. Annotators were then asked to evaluate the models’ outputs for each task using the metrics we developed, which include *accuracy*, *clarity*, *helpfulness*, *appropriateness*, *correctness*, *consistency*, *depth* of explanation, *quality* of text refinement, *fluency*, and *relevance*. Each response was scored on a scale from 1 to 5, reflecting on how well it met the evaluation criteria. Detailed descriptions of the evaluation scales are provided in Appendix G.

We calculate the Cohen’s Kappa scores for inter-annotator agreement among four annotators across all tasks. Figure 3 displays the overall inter-annotator agreement, which varies across annotator pairs, with some pairs showing moderate agreement and others showing slight to fair agreement. Individual inter-annotator agreements are provided in Appendix G.1.

5.2 Results

We present the results of our evaluation in Figure 4. The performance of each model on the outlined writing tasks is discussed below.

GEC and Explanation Tasks. For the GEC and explanation tasks, we evaluate the performance of the models based on three criteria: *accuracy*, *clarity*, and *helpfulness*. Figure 4a displays the evaluation result for this task.

Overall, GPT-4o demonstrates superior performance in the GEC task, excelling in all criteria. Cohere Command R+ and Gemini 1.5 Pro provide balanced results, whereas GPT-4, while maintaining a reasonable level of *clarity*, falls short in *accuracy* and *helpfulness*.

MWEs. In the MWEs task, we evaluate the four models based on *clarity*, *appropriateness* of multi-word expressions, and *relevance*. Figure 4b shows result of the evaluation for this task. Overall, GPT-4o outperforms all models in the MWEs task, delivering the highest scores in all criteria. Both Cohere Command R+ and Gemini 1.5 Pro perform well, particularly in *clarity* and *relevance*, while GPT-4 lags in *appropriateness* but maintains reasonable levels of *clarity* and *relevance*.

Text Refinement. We evaluate the text refinement task based on *clarity*, *quality* of text refinement, and *fluency*. Figure 4c presents result of the evaluation for this task. GPT-4o excels in the Text Refinement task, leading with superior scores in *clarity*, *quality* of text refinement, and *fluency*. Both Cohere Command R+ and Gemini 1.5 Pro show balanced and consistent performance, while GPT-4 also performs well, particularly in *fluency*.

Grammatical Rules and Definitions. For this task, we evaluate the models based on three criteria: *clarity*, *consistency* with grammatical rules and definitions, and *depth* of explanation. Figure 4d illustrates the result for this task.

In summary, GPT-4o performs exceptionally in the grammatical rules and definitions task, scoring highest in all criteria. Gemini 1.5 Pro follows closely, offering strong and balanced performance. Cohere Command R+ also performs well, particularly in *clarity* and *depth*, while GPT-4 shows the lowest scores across all criteria.

I’rab. We evaluate the models on three criteria for the I’rab task: *clarity*, *helpfulness*, and *correctness* of I’rab. Figure 4e presents result of the evaluation for this task.

Overall, GPT-4o dominates the I’rab task with superior scores in *clarity*, *helpfulness*, and *correctness*. Cohere Command R+ also performs well, especially in *clarity* and *helpfulness*. Gemini 1.5 Pro shows balanced results, while GPT-4 has the lowest scores across all criteria.

5.3 Discussion

Cohere Command R+. Cohere Command R+ excels in delivering clear and helpful explanations,

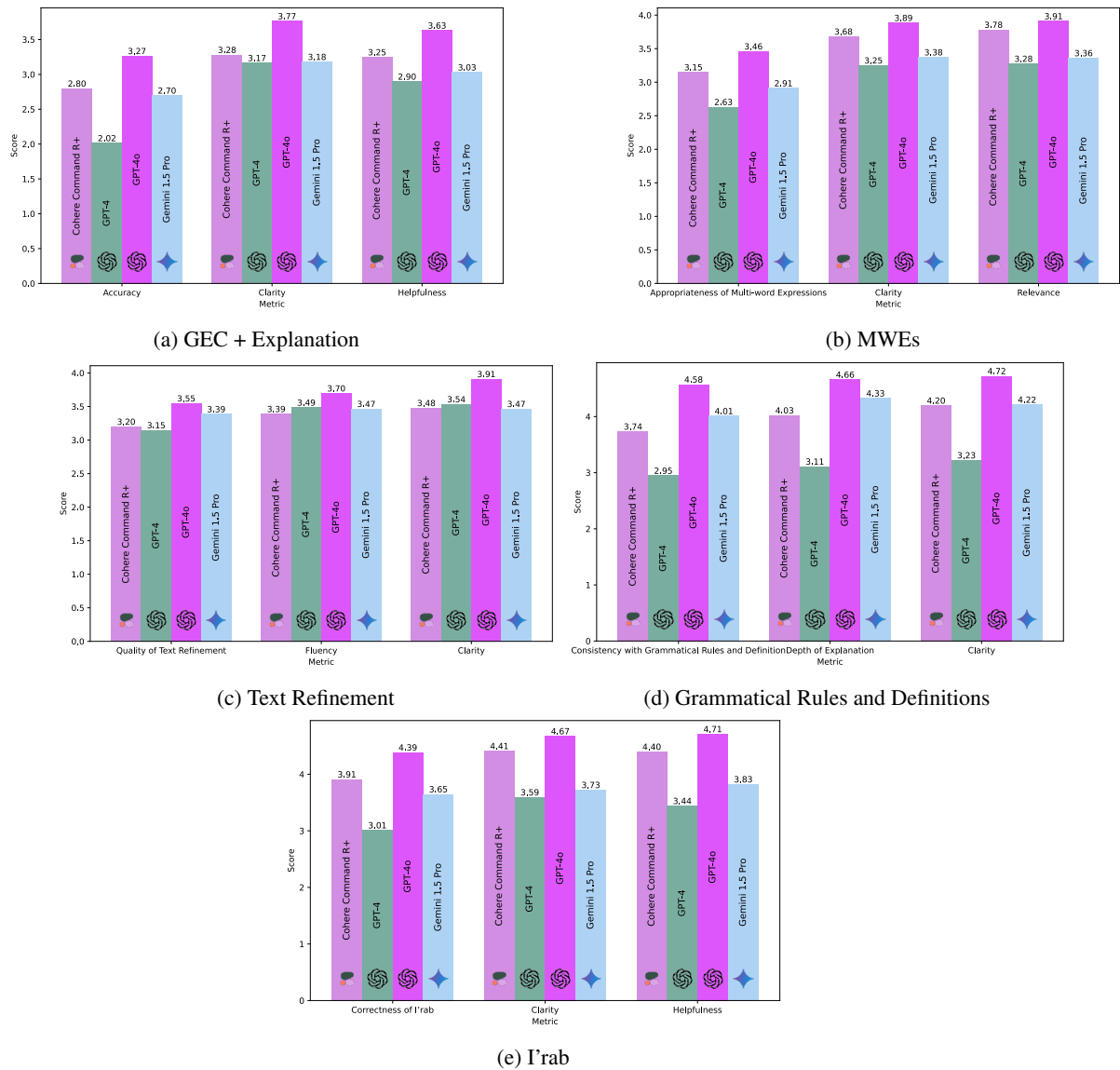


Figure 4: Results of human evaluation for four LLM models: GPT-4, GPT-4o, Cohere Command R+, and Gemini 1.5 Pro on five subtasks in *Gazelle*.

particularly in handling MWEs and I'rab writing tasks. It consistently offers understandable and relevant responses in these areas. However, its performance in GEC tasks is less robust, with lower *accuracy* and *correctness* compared to other models. We provide detailed statistical evaluation metrics for the five subtasks across all models in Table G.3

GPT-4. GPT-4 demonstrates the weakest overall performance among all the models evaluated. While it does provide reasonable *clarity* and *relevance* in its responses, it struggles with the *accuracy* and *depth* required for tasks such as GEC, explaining grammatical rules, and providing GEC explanation. It scores the lowest compared to the

other models in these criteria. Although GPT-4's performance is stable, it does not excel in any specific task.

GPT-4o. GPT-4o stands out as the top-performing model across all tasks. It consistently achieves the highest scores by demonstrating superior *clarity*, *correctness*, and *depth* in its responses. GPT-4o reliably provides the most *accurate*, *relevant*, and *helpful* outputs, making it the most effective model for comprehensive Arabic writing assistance. Its performance is robust and reliable, with minimal weaknesses observed.

Gemini 1.5 Pro. Gemini 1.5 Pro delivers a balanced and strong performance across all tasks.

It particularly excels in explaining grammatical rules and providing detailed and consistent responses. Although it performs well overall, it does not quite reach the high performance levels of GPT-4o. In tasks involving MWEs and I'rab, Gemini 1.5 Pro shows room for improvement, where it is less effective compared to GPT-4o.

5.4 Human Analysis of the Models

We outline insights from our human analysis, offering a detailed evaluation of how each model handles specific linguistic challenges across all tasks.

GEC and explanation task. Models frequently corrected sentences but occasionally struggled with accurate explanations or identifying all errors, highlighting their limitations in grasping Arabic grammar complexities. The task of applying grammatical rules and definitions revealed varying proficiency among the models, showcasing their differing capabilities in handling Arabic grammar.

MWEs and Metaphors. All models struggled with handling MWEs and metaphors. This difficulty was evident in models' ability to identify errors and provide accurate metaphors or MWEs, highlighting their inability to fully capture the linguistic richness of Arabic.

The complexity of handling MWEs and metaphors in Arabic stems from their nuanced and context-dependent nature. Unlike single-word expressions, MWEs and metaphors often convey meanings that cannot be directly inferred from the individual words. Furthermore, the cultural and contextual connotations embedded in Arabic expressions add another layer of difficulty for the models.

Text Refinement. In the text up-refinement task, a creative exercise, models showed diverse capabilities. Some excelled in enhancing stylistic elements such as GPT-4o and Gemini 1.5 Pro, while others focused on basic corrections as in the case of GPT-4.

I'rab. The I'rab task, requiring deep linguistic knowledge, revealed mixed abilities in applying these rules. Models like Cohere Command R+ and GPT-4o performed well, while others showed limitations in providing a fully correct I'rab of the sentences.

Overall, this analysis highlights the variability in model performance, showcasing each model's strengths in specific areas and their limitations in understanding and applying Arabic grammar and stylistic enhancements.

6 Conclusion

In this work, we present *Gazelle* a comprehensive dataset and evaluation framework to enhance Arabic writing assistance tools. Our manually fine-grained dataset incorporates five writing subtasks under two main themes. Our analysis uncovers both strengths and limitations of the most leading LLMs in tackling the unique challenges of Arabic writing. These insights highlight the need for continuous model training and dataset enrichment to address the complexities of Arabic language processing.

Our study showcases the varied capabilities of models in handling Arabic grammar and stylistic nuances, providing a foundational step towards more advanced 'Human-AI co-writing' tools for Arabic writing. Future research may include on expanding the dataset and integrating diverse linguistic features to develop robust and reliable models for mastering Arabic writing.

Overall, our work serves as a stepping stone towards creating more effective Arabic writing assistance tools and advancing the field significantly.

Limitations

We identify the following limitations in this work:

- The dataset may not encompass the full diversity of the Arabic language, including its many dialects and regional variations, limiting the generalizability of the results.
- The study focuses on zero-shot settings, which highlight models' inherent capabilities but not their performance with fine-tuning. Future work may include fine-tuning for enhanced performance.
- The human evaluation process is subjective and may be influenced by individual biases, particularly in tasks like text up-scaling and error explanation.
- The study evaluates four specific LLMs. As models evolve, re-evaluation will be necessary to maintain relevance in assessing their performance on Arabic writing tasks.

Ethics Statement

Encouraging research development and contributing to a collaborative research culture.

Low-resource languages such as Arabic face significant hurdles in the development of advanced AI tools due to a scarcity of comprehensive datasets. This shortage hampers the ability to train effective models, limiting the creation of sophisticated writing assistance tools for these languages. Our current work aims to bridge this gap by introducing *Gazelle*, a dataset specifically curated for Arabic writing. This dataset is designed to support the development of AI-powered tools that can enhance Arabic language processing. By providing this diverse and detailed dataset, we aspire to stimulate further research and development in Arabic AI tools, fostering a collaborative effort to overcome the challenges faced by low-resource languages.

Advancing Pedagogical Approaches in Arabic Writing. With the growing interest in enhancing language proficiency, the accuracy and effectiveness of written communication have become essential for both native speakers and second language learners. LLMs are increasingly utilized as writing assistants, highlighting their significant role in educational tools. *Gazelle*, supports the development of AI-driven tools that facilitate learning and teaching in Arabic writing. *Gazelle* focuses on a broad spectrum of writing tasks and aims to assist both non-native Arabic learners and native speakers in understanding the pedagogical aspects of Arabic writing. This approach not only bridges the gap between learners and fluent written communication but also enriches the educational tools available for teaching Arabic. By offering comprehensive resources tailored to diverse writing needs, *Gazelle* enhances the capability of educational platforms to help users master the intricacies of Arabic writing.

Data privacy. In relation to the data used in this work, all datasets are publicly available. Therefore, we do not have privacy concerns.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Al-

liance of Canada,⁵ and UBC Advanced Research Computing-Sockeye.⁶

References

- Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. [NADI 2020: The first nuanced Arabic dialect identification shared task](#). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Aman Agarwal. 2022. [Explain to me like i am five-sentence simplification using transformers](#). *arXiv preprint arXiv:2212.04595*.
- Muhammed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfasasi. 2018. [A leveled reading corpus of Modern Standard Arabic](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Molham Al-Maleh and Said Desouki. 2020. [Arabic text summarization using deep learning approach](#). *Journal of Big Data*, 7.
- Abdullah Alfaifi and ES Atwell. 2014. [An evaluation of the arabic error tagset v2](#). In *Proceedings of the ACL 2014-The American Association for Corpus Linguistics conference*. The American Association for Corpus Linguistics.
- Bashar Alhafni, Reem Hazim, Juan Piñeros Liberato, Muhammed Al Khalil, and Nizar Habash. 2024. [The samer arabic text simplification corpus](#). *arXiv preprint arXiv:2404.18615*.
- Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. [Advancements in arabic grammatical error detection and correction: An empirical investigation](#). *arXiv preprint arXiv:2305.14734*.
- Riadh Belkebir and Nizar Habash. 2021a. [Automatic error type annotation for arabic](#). *arXiv preprint arXiv:2109.08068*.
- Riadh Belkebir and Nizar Habash. 2021b. [Automatic error type annotation for arabic](#). In *Conference on Computational Natural Language Learning*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

⁵<https://alliancecan.ca>

⁶<https://arc.ubc.ca/ubc-arc-sockeye>

- Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017a. [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.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017b. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *arXiv preprint arXiv:2303.12712*.
- Eric Chamoun, Michael Schlichktrull, and Andreas Vlachos. 2024. [Automated focused feedback generation for scientific writing assistance](#). *arXiv preprint arXiv:2405.20477*.
- Yanhan Chen, Hanxuan Wang, Kaiwen Yu, and Ruoshui Zhou. 2024. [Artificial intelligence methods in natural language processing: A comprehensive review](#). *Highlights in Science, Engineering and Technology*, 85:545–550.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. [Palm: Scaling language modeling with pathways](#). *Journal of Machine Learning Research*, 24(240):1–113.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53.
- Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. [Understanding iterative revision from human-written text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.
- Yasmin Einieh and Amal AlMansour. 2022. [Deep learning in arabic text summarization: Approaches, datasets, and evaluation metrics](#). In *2022 20th International Conference on Language Engineering (ESOLEC)*, volume 20, pages 45–49.
- Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. [Enhancing grammatical error correction systems with explanations](#). *ArXiv*, abs/2305.15676.
- Charles A. Ferguson. 2020. [Diglossia](#). *The Bilingualism Reader*.
- Linda Flower and John R. Hayes. 1980. [The cognition of discovery: Defining a rhetorical problem](#). *College Composition and Communication*, 31(1):21–32.
- 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.
- Abdenbi Hamani. 2014. [The phenomenon of i’rab: Between linguistic conventions and relations of proximity](#). *Tabayyun*, 2(7):47–70.
- Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. [Arabic word-level readability visualization for assisted text simplification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.
- Clive Holes. 2004. [Modern Arabic: Structures, functions, and varieties](#). Georgetown University Press.
- Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. [Creative writing with an ai-powered writing assistant: Perspectives from professional writers](#). *arXiv preprint arXiv:2211.05030*.
- Jenny Kunz and Marco Kuhlmann. 2024. [Properties and challenges of llm-generated explanations](#). *arXiv preprint arXiv:2402.10532*.
- Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [Beyond english: Evaluating llms for arabic grammatical error correction](#). *arXiv preprint arXiv:2312.08400*.
- Samira Lagrini and Mohammed Redjimi. 2021. [A new approach for arabic text summarization](#). In *International Conference on Natural Language and Speech Processing*.
- Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambansgans, David Zhou, Emad A Alghamdi, et al. 2024. [A design space for intelligent and interactive writing assistants](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–35.
- Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. [The value, benefits, and concerns of generative ai-powered assistance in writing](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. [Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning](#). *Advances in Neural Information Processing Systems*, 35:1950–1965.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. [The penn arabic treebank: Building a large-scale annotated arabic corpus](#). In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.
- Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Belguith Hadrach. 2022. [Annotating verbal multiword expressions in arabic: Assessing the validity of a multilingual annotation procedure](#). In *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1839–1848.
- Djamel Mostefa, Jaber Abualasal, Omar Asbayou, Mahmoud Gzawi, and Ramzi Abbès. 2015. [Techlimed@qalb-shared task 2015: a hybrid arabic error correction system](#). In *ANLP@ACL*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. [Instruction tuning with gpt-4](#). *arXiv preprint arXiv:2304.03277*.
- Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. [Coedit: Text editing by task-specific instruction tuning](#). *arXiv preprint arXiv:2305.09857*.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghoulani, 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.
- Alla Rozovskaya, Nizar Habash, Ramy Eskander, Noura Farra, and Wael Salloum. 2014. [The columbia system in the qalb-2014 shared task on arabic error correction](#). In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 160–164.
- Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil. 2018. [Feature optimization for predicting readability of Arabic L1 and L2](#). In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29, Melbourne, Australia. Association for Computational Linguistics.
- Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. [Multiword expressions: A pain in the neck for nlp](#). In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. [Multitask prompted training enables zero-shot task generalization](#). *arXiv preprint arXiv:2110.08207*.
- Majdi Sawalha, Eric Atwell, and Mohammad Abd-Alrahman Mahmoud Abushariah. 2013. [Salma: Standard arabic language morphological analysis](#). *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.
- Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. [Peer: A collaborative language model](#). *arXiv preprint arXiv:2208.11663*.
- Aiman Solyman, Zhenyu Wang, and Qian Tao. 2019. [Proposed model for arabic grammar error correction based on convolutional neural network](#). *2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–6.
- Aiman Solyman, Wang Zhenyu, Tao Qian, Arafat Abdulgader Mohammed Elhag, Muhammad Toseef, and Zeinab Aleibeid. 2021. [Synthetic data with neural machine translation for automatic correction in arabic grammar](#). *Egyptian Informatics Journal*, 22(3):303–315.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. [Gee! grammar error explanation with large language models](#). *arXiv preprint arXiv:2311.09517*.
- Alex Urmeneta and Margarida Romero. 2024. *Creative Applications of Artificial Intelligence in Education*.
- Kees Versteegh. 2014. *Arabic language*. Edinburgh University Press.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.

- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. [A prompt pattern catalog to enhance prompt engineering with chatgpt](#). *arXiv preprint arXiv:2302.11382*.
- Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Os-sama 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).
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. [Instruction tuning for large language models: A survey](#). *arXiv preprint arXiv:2308.10792*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. [Opt: Open pre-trained transformer language models](#). *arXiv preprint arXiv:2205.01068*.

Appendices

We provide an addition organized as follows:

- Detailed Related Works [A](#).
- Linguistic Background of Arabic [B](#).
- Extended ALC Error Taxonomy [D](#).
- ATB Synthetic Data [E](#).
- Prompt Design [F](#).
- Evaluation matrices for Arabic Writing Assistant Tasks [G](#).

A Detailed Related Works

LLM’s for Writing assistance. Advancements in instruction tuning for writing assistance include works like PEER ([Schick et al., 2022](#)), which enhances editing by adhering to and explaining a user-provided plan, and CoEdIT ([Raheja et al., 2023](#)), which focuses on managing editing plans to refine and clarify content without introducing new information. Additionally, SWIF²T ([Chamoun et al., 2024](#)), designed for scientific writing, generates specific, actionable comments to identify weaknesses in a paper and suggest revisions. These developments highlight the progress in creating sophisticated tools that enhance the writing and editing processes across various applications.

Progress in Arabic Writing Tasks. In GEC, advances include the QALB corpus used for the QALB 2014 and 2015 shared tasks ([Zaghouani et al., 2014](#); [Rozovskaya et al., 2015](#)), and the Zaebuc corpus ([Habash and Palfreyman, 2022](#)). Advances in model techniques include Convolutional Neural Networks, alignment algorithms, LLMs ([Solyman et al., 2019, 2021](#); [Alhafni et al., 2023](#); [Kwon et al., 2023](#)). In text simplification, tools such as visualization for assisted text simplification ([Hazim et al., 2022](#)) and feature optimization techniques ([Saddiki et al., 2018](#)) have been developed, along with corpus advancements like the SAMER corpus ([Al Khalil et al., 2018](#); [Alhafni et al., 2024](#)). For text summarization, model developments include survey approaches ([Einieh and AIMansour, 2022](#)), rhetorical structure theory with statistical methods ([Lagrini and Redjimi, 2021](#)), and deep learning approaches ([Al-Maleh and Desouki, 2020](#)).

Explainable Writing Systems. Explainable writing systems enhance the transparency of language processing tools by providing clear and understandable insights aligned with instructional goals. For example, ERRANT ([Bryant et al., 2017b](#)) annotates grammatical errors with clear edit boundaries and error types. EXPECT ([Fei et al., 2023](#)) offers a dataset with detailed annotations to support explainable grammatical error correction. GEE ([Song et al., 2023](#)) advances this by providing one-sentence explanations for grammatical errors in pairs of erroneous and corrected sentences. In text simplification and feedback generation, systems strive to provide clear and relevant explanations ([Agarwal, 2022](#)). These efforts highlight the importance of creating tools that deliver understandable insights and ensure alignment with instructional goals. Additionally, understanding the properties and challenges of LLM-generated explanations ([Kunz and Kuhlmann, 2024](#)) is crucial for maintaining user trust and comprehension.

B Linguistic Background of Arabic

Arabic is a highly sophisticated and complex language, distinguished by the richness of its morphological and grammatical features ([Holes, 2004](#)). It is linguistically diverse, encompassing numerous dialects across a widespread geographical region ([Versteegh, 2014](#)). Arabic is recognized as a diglossic language, where both Modern Standard Arabic (MSA) and various dialects coexist ([Ferguson, 2020](#)). MSA is employed in formal contexts, whereas local dialects dominate everyday communication, exhibiting significant differences in pronunciation, vocabulary, and grammar. The inherent complexity and diversity of Arabic presents considerable challenges in the field of NLP, making efforts in writing tasks and writing assistance valuable. For instance, Arabic verbs are intricate, changing based on tense, mood, voice, and gender, posing significant challenges for writing tasks. Moreover, the grammar includes complex rules, such as changes in noun forms depending on their grammatical role. Additionally, Arabic is orthographically ambiguous due to its reliance on consonants and the optional use of diacritics to denote short vowels and other phonetic details ([Sawalha et al., 2013](#)). This ambiguity leads to multiple interpretations of the same consonant string without context or diacritics, complicating writing in Arabic ([Abdul-Mageed et al., 2020](#); [Belkebir and](#)

Habash, 2021a). For example, the word علم with diacritics has three different meanings: عَلِمَ meaning (he knew), عِلْمٌ meaning (knowledge), and عَلَمٌ meaning (flag).

Furthermore, the morphological ambiguity of Arabic occurs when the same root or pattern leads to different meanings based on context, inflection, or derivation. Arabic's rich system of roots and patterns (templates) allows for a single root to generate various words with different meanings and grammatical roles. The root of the verb ض-ر-ب (hit), for instance, has different patterns indicating tense, voice (active/passive), and other grammatical nuances, such as ضَرَبَ (daraba - he hit) vs. يُضْرَبُ (yu'arib - he is hitting).

C Writing Assistance Instructions

GEC Instructions. For example, the Orthographic error class features a subclass for *hamza* errors, further divided into four sub-subclasses: *medial hamza*, *final hamza*, *cutting hamza*, and *hamzet al-wasl*. Our instructions address a broad spectrum of these sub-classes and sub-subclasses, ensuring comprehensive coverage across different error categories.

GEC Error Explanations. In our dataset, we provide explanation of the GEC errors in both Arabic and English. For instance, consider the sentence يجري ضد الحقوق الإنسان (goes against human rights), which contains a *Syntax* error. The English explanation is: Adding the definite article ال to the annexed noun is incorrect because the annexed noun and the annexing noun should share the same definiteness or indefiniteness. In another example, شعرت بالسعادة عندما وطأت أرض الوطن, which contains an *Orthographic* error, the Arabic explanation is: وطئت: كتبت الهمزة المتوسطة على ياء لأن حركتها السكون وحركة الحرف الذي قبلها الكسرة، والكسرة أقوى من السكون وتناسبها الياء، لذلك كتبت على ياء.

ATB Synthetic Data. We utilize all parts of the ATB to generate 100 examples related to different classes of errors in our taxonomy, covering the following error classes and sub-classes:

- **Orthographic** errors, including:

i) *Wrong Order of Word Characters (OC)*, ii) *Replacement in Word Characters (OR)*, iii) *Additional Characters (OD)*, iv) *Missing Characters (OM)*.

- **Syntax** errors, which include:

i) *Missing Words (XM)*, ii) *Unnecessary Words (XT)*.

- **Morphological** errors, involving: i) *Word Inflection (MI)*, ii) *Verb Tense (MT)*.

- **Split and Merge**

- **Punctuation** errors, categorized into:

i) *Punctuation Confusion (PC)*, ii) *Missing Punctuation (PM)*, iii) *Unnecessary Punctuation (PT)*, iv) *Other Punctuation Errors (PO)*.

- **Text Refinement**

For instance, the informal use of the following expression in MSA: نعم، يا سعد بيه، ليس فيها أي خدوش، would be polished to be: نعم، يا سيد سعد، ليس فيها أي خدوش. The explanation provided is: The word بيه is an Egyptian title indicating respect and does not exist in Standard Arabic, so it was necessary to phrase it as سيد. Another example of text refinement is أنتم ما زلتم جالسن ترغون أما زلتما تثرثران؟. The explanation provided: أما ترغون is an incorrect word and does not exist in Standard Arabic; the correct equivalent is اثرثرة.

D Extended ALC Error Taxonomy

Orthographic Errors refer to mistakes in the spelling of a word. It occurs when the arrangement of letters does not match the standard or accepted spelling. These errors can happen for several reasons, such as typing mistakes, incorrect memory of the spelling, or unfamiliarity with the spelling rules of a language. Orthographic errors are particularly common in languages with complex and irregular spelling rules, like Arabic. For instance, there are common spelling errors in *medial hamza* as in راي instead of رأي. Other error examples include the confusion between *Ta mabsouta* and *Ta marbouta* as in طارة instead of طارت. Within the Orthographic errors category, which originally encompasses 12 subclasses as per the ALC framework, we have introduced additional granularity by defining further sub-subclasses. These include four distinct categories for Hamza errors: i) *Medial hamza*, ii) *Final hamza*, iii) *Cutting hamza*, and iv) *Hamzat*

al-wasl. Additionally, we address the common confusions between Ha, Ta, and Ta marbouta by subdividing them into i) *Ta mabsouta*, ii) *Ta marbouta*, and iii) *Ha*.

Syntax Errors in language use refer to errors in the structure of a sentence which include the arrangement of words and phrases in ways that do not conform to the grammatical rules of the language, making the sentence unclear or incorrect. Syntax errors can involve incorrect word order, missing components like subjects or verbs, or improper use of grammatical forms such as agreement, or conjugation. In Arabic, errors include the gender, number, cases, etc. For example, *أطلق الجندي رصاصتان* the error in dual word *رصاصتان* which should be *صاصتين* in Arabic. For the Syntax error category, which initially comprised seven subclasses, we have introduced an additional subclass titled *Closed class errors*. Further, the gender error subclass is now refined into a sub-subclass errors of using feminine and masculine forms and vice versa. The number subclass has been divided into four sub-subclasses: i) *singular*, ii) *dual*, iii) *plural*, and iv) *numbers* with additional subdivisions under plural—namely i) *sound masculine plural*, ii) *sound feminine plural*, iii) *broken plural*, and iv) *Al-maqsour plural*. Additionally, we have structured the number into three further sub-subclasses: i) *singular numbers*, ii) *composite numbers*, and iii) *Al-Aqoud*.

We further refined the taxonomy by introducing two sub-subclasses under definiteness, specifically addition and omission of the definite article "Al." Additionally, we have delineated five sub-subclasses within the category of case errors: i) *nominative*, ii) *genitive*, iii) *accusative*, iv) *jussive*, v) and *other case errors*. We also introduced a new subclass titled *Closed class category*. This category captures miscellaneous grammatical inaccuracies involving i) *pronouns*, ii) *the five verbs*, and iii) *the five nouns* that do not fit neatly into other predefined categories. The pronouns sub-subclass is divided into three specific issues: i) *adding unnecessary pronoun*, ii) *removing necessary pronoun*, iii) and *wrong pronoun reference*.

Semantic Errors refer to an error in the meaning or interpretation of a sentence or phrase, despite it being grammatically correct. This type

of error occurs when words or phrases are used inappropriately, leading to confusion or a failure to convey the intended message. Semantic errors can stem from using a word that does not have the intended meaning of the sentence, or from a logical inconsistency in the statement. For instance, adding unnecessary conjunction as *قرات الكتاب والذي كان مفيداً*, the error here lies in the extra conjunction *واو* in the sentence. Also, there is an error of missing the preposition *إلى* in *محاولة المنازل ثكنات عسكرية*. In the semantics category, which initially comprised three subclasses, we added a new subclass designated *Special Expressions*. This subclass addresses the incorrect usage of special expressions at either the word or phrase level. Breakdown of this category includes a Word Selection Error subclass divided into five further categories: i) *Adding unnecessary preposition*, ii) *Removing necessary preposition*, iii) *Replacing preposition*, iv) *Preposition use*, v) and *other word selection errors*. Additionally, for conjunction errors, we organized this into four sub-subclasses: i) *Adding unnecessary conjunction*, ii) *Removing necessary conjunction*, iii) *Replacing conjunction*, iv) and *other Atf errors*. The subsubclasses we introduced in the semantic errors might appear to overlap with the missing and additional characters in the syntax class, but the errors we generated impact the sentence's semantic meaning rather than its structural integrity.

Punctuation Errors occur when punctuation marks are misused, omitted, or placed incorrectly within sentences. These errors can alter the clarity, flow, and meaning of text. Typical punctuation errors include the misuse of commas, periods, semicolons, colons, and quotation marks, etc. For instance, the confusion of using question mark in *ما أروع القمر؟* instead of using the exclamation mark *ما أروع القمر!*. Within the framework of the original error taxonomy, punctuation errors are categorized into four subclasses: i) *missing punctuation*, ii) *punctuation confusion*, iii) *unnecessary punctuation*, iv) and *other punctuation errors*. These subclasses are further divided into detailed sub subclasses that specifically address various punctuation marks, such as commas, exclamation marks, and full stops, etc.

Arabic Definition	ما جمع التكسير؟	<p>الإجابة: هو ما دلّ على أكثر من اثنين، وسمي بـ(جمع التكسير) لأن بناء المفرد يغيّر (ويتكسر) عند الجمع، على نحو: كتاب-كتب. عالم-علماء.</p> <p>ويكون التغيير على بنية الكلام إما بزيادة أو نقصان بعدد الحروف، وجمع التكسير يأتي على نوعين؛ جمع قلة وجمع كثرة.</p>
Definition	Explain the concept of " hollow verbs) in Arabic grammar and provide examples.	<p>الأفعال الجوفاء (hollow verbs) are verbs with a weak letter و or ي in the second position of the base letters.</p> <p>Examples include:</p> <ul style="list-style-type: none"> • "ق، و، ل" he said from base letters "قال" • "ب، ي، ع" he sold from base letters "باع" <p>These verbs undergo specific morphophonemic changes to avoid weakness and ensure correct pronunciation.</p>
Arabic GEC	<p>سأعطيك جملة بها خطأ نحوي، حدد نوع الخطأ وقم بتصحيحه مع شرح السبب لهذا الخطأ.</p> <p>الجملة: سيدي الرئيس نلتمس منكم وأنتم الذين تعملون جاهدين على إرساء جزائر جديدة.</p>	<p>الإجابة: بكل تأكيد، سأقوم بفحص الجملة وتحديد نوع الخطأ وشرحه.</p> <p>نوع الخطأ: جمع المذكر السالم.</p> <p>الجملة المصححة: سيدي الرئيس نلتمس منكم وأنتم الذين تعملون جاهدين على إرساء جزائر جديدة.</p> <p>شرح الخطأ: خطأ في علامة جمع المذكر السالم لأن جاهدين تجمع جمع منكر سالم، وجمع المذكر السالم ينصب بالياء والنون.</p>
GEC	<p>Analyze the sentence and identify any errors in the use of numbers. Make the necessary edits to correct the error and ensure that the sentence is grammatically correct.</p> <p>The sentence: وضرورة التخلص من كثير من العادات والسلوكيات</p>	<p>I will review the error and make appropriate adjustments.</p> <p>The error: There is an error in this word: والسلوكيات, where it should be: والسلوكات.</p> <p>Explanation: The correct form is "مُشْتَرِكَات" because "مُشْتَرِكٌ" is a passive participle and exceeds four letters, and when made dual, its alif is changed to ya.</p> <p>The correct sentence: وضرورة التخلص من كثير من العادات والسلوكات</p>
Iraab	اعرب الجملة التالية إعراباً كاملاً: الجملة: راقب الوالد أولاده وهم يلعبون.	<p>بكل تأكيد، إليك إعراب الجملة بالكامل:</p> <p>راقب: فعل ماضٍ مبني على الفتح.</p> <p>الوالد: فاعل مرفوع وعلامة رفعه الضمة مبني في محل جر مضاف إليه.</p> <p>الواو: واو الحال (هم) ضمير رفع منفصل مبني في محل رفع مبتدأ.</p> <p>يلعبون: فعل مضارع مرفوع وعلامة رفعه ثبوت نون الإعراب في آخره خبر المبتدأ (هم) والجملة الاسمية من (هم يلعبون) في محل نصب حال.</p>

Figure D.1: Examples of writing assistance instructions in Arabic and English.

Class	Tag	Subclass	Sub2	Sub3	Description	Example	
Orthography	OH	Hamza	Medial Hamza	—	Medial Hamza E.	كتائب ← كتائب	
			Final Hamza	—	Final Hamza E.	تباطو ← تباطؤ	
			Cutting Hamzah	—	Cutting Hamzah E.	اسهمت ← أسهمت	
			Hamzet al-Wasl	—	Hamzet al Wasel E.	استعلمت ← استعلمت	
	OA	C. in Alif Ya and Alif-Maqsura	Shortening Alif/ Alif Maqsoura	—	C. in Alif Ya, and Alif-Maqsura	مستوي ← مستوى	
	OW	C. in Alif Fariqa	—	—	—	سموا ← سمو	
			Ta Mabsouta	—	C. in Ta Mabsouta	طارة ← طارت	
	OT	C. in Ha, Ta and Ta-Marbuta	Ta Marbuta Ha	—	C. in Ta Marbuta C. in Ha	أتنزه ← أتنزه كتابه ← كتابه	
	ON	C. between Nun and Tanwin	—	—	C. between Nun and Tanwin	منايرن ← مناير	
	OS	Shortening long vowels	—	—	Shortening long vowels	ذكرتهم ← ذكروهم	
	OG	Lengthening short vowels	—	—	Lengthening short vowels	هاكنا ← هكنا	
	OC	Wrong order of word Chr.(s)	—	—	Wrong order of word Chr.(s)	كعب ← بك	
	OR	Replacement in Word Chr.(s)	—	—	Rep. in word Chr.(s)	تحويل ← تحوّل	
	OD	Additional Chr.(s)	—	—	Adding unnecessary Chr.(s)	اطسم ← اسم	
	OM	Missing Chr.(s)	—	—	Missing necessary Chr.(s)	قنال ← قنابل	
	OO	Other Orthographic E.	—	—	Other orthographic E.	—	
	Syntax	XG	Gender	Fem. Vs. Masc.	—	C. in Fem. Vs. Masc.	يشرب ← تشرب
				Singular	—	C. in Sing. Vs. Pl.	النحل ← النحلة
Dual				—	C. in Dual	المعلمين ← المعلمان	
XN		Number	Plural	Sound Masc. Pl.	Wrong Sound Masc. Pl.	مؤذب ← مؤدبون	
				Broken Pl.	Wrong Broken Pl.	سعيدين ← سعءاء	
			Almaqsour Pl.	Wrong Almaqsour Pl.	المشتروات ← المشتریات		
			Numbers	Sound Fem. Pl.	Wrong Sound Fem. Pl.	كتبوا ← كتبن	
XF		Definiteness	—	Sing. Num.	C. in Sing. Num.	ثلاث أيام ← ثلاثة أيام	
				Composite Num.	C. in Composite Num.	أربع عشر مبدا ← أربعة عشر	
				Al-Aqoud	C. in Al-Aqoud	الخمسون ← الخمسين	
				Adding "Al"	—	Adding "Al"	المنتصف ← منتصف
XC		Case	—	Removing "Al"	—	C. in Removing Al	عالمي ← العالمي
				Nom. Case E.	—	C. in Nom. Case	الممرضين ← الممرضون
				Gen. Case E.	—	C. in Gen. Case	على عرفاتاً ← عرفات
				Acc. Case E.	—	C. in Acc. Case	عادي ← عاديا
XR		Closed Class E.	—	Juss. Case E.	—	C. in Juss. Case	يدعوا ← يدع
				Other Case E.	—	Other Case E.	—
				Pronouns	Adding Unnecessary Pron.	C. in Adding Pron.	انا رأيته ← رايته
	Removing Necessary Pron.				C. in Removing Pron.	أخذنا ← أخذنا	
XM	Missing Word	—	Wrong Pron. Ref.	C. in Pron. Ref.	منازلهم ← منازلهن		
			The Five N.	—	C. in Five N.	أبوك ← أبيك	
			The Five V.	—	C. in Five V.	تقنعى ← تقنعين	
			—	—	Missing Certain word	تستضيف ← ان تستضيف	
XT	Unnecessary Word	—	—	Adding Unnecessary Word	حق الحالة ← الحالة		
XO	Other Syntactic E.	—	—	Other Syntactic E.	—		
Semantics	SW	Word Selection E.	Removing Necessary Prep.	—	C. in Removing Necessary Prep.	الربيع ← للربيع	
			Replacing Prep.	—	C. in Replacing Prep.	في سيارتنا ← بسيارتنا	
			Adding Unnecessary Prep.	—	C. in Adding Unnecessary Prep.	من وراء ← وراء	
			Other Preposition E.	—	—	—	
	SF	Conjunction E.	—	Preposition Use	—	C. in Prep. Use	كطريقة ← طريقة
				Other Word Selection E.	—	C. in Other Word Selection E.	—
				Adding Unnecessary Conj.	—	C. in Adding Unnecessary Conj.	حتى إلى ← إلى
				Removing Necessary Conj.	—	C. in Removing Necessary Conjunction	المدير، الموظفون ← المدير والموظفون
	SS	Special Expression	—	Replacing Conj.	—	C. in Replacing Conj.	فالمجتمع ← والمجتمع
				Other Atf E.	—	C. in Other Atf E.	مع مديره ← ومديره
SO	Other Semantic E.	—	—	C. in Special Expression	هامة ← مهمة		
Morph	MI	Word Inflection	—	—	Other Semantic E.	—	
			—	—	C. in Word Inflection	مشينة ← شائنة	
			—	—	C. in Verb Tense	يعيش ← عاش	
			—	—	Other Morphological E.	—	
Punctuation	PC	Punctuation C.	Exclamation Mark (!)	—	E. in Exclamation Marks.	ما أجمل! ← ما أجمل!	
			Semicolon (;)	—	E. in Semicolon.	ذهبت إلى الطبيب؛ ← ذهبت إلى الطبيب	
			Colon (:)	—	E. in Colon.	قال المعلم: ← قال المعلم:	
			Ellipsis (...)	—	E. in Ellipsis.	فيقول أحدهما. ← فيقول أحدهما.	
			Quotations Marks (" ")	—	E. in Quotation Marks.	—	
			Square Brackets []	—	E. in Square Brackets.	نعم! ← نعم	
			Comma (,)	—	E. in Comma.	الشعر، ← الشعر	
			Question Mark (?)	—	E. in the Question Marks.	كيف ذهبت. ← كيف ذهبت؟	
			Parenthesis ()	—	E. in Parenthesis.	مؤلفة (من) ← مؤلفة من	
			Dash (-)	—	E. in the Dash.	بعض أفرادها ← بعض أفرادها	
Period/full stop (.)	—	E. in full stop.	يجب أن تقوم انتفاضة ← يجب أن تقوم انتفاضة				
Merge	MG	—	—	Words are merged	إنشاء الله ← إن شاء الله		
Split	SP	—	—	Words are split	لولا ← لولا		

Table D.1: The ALC error taxonomy has been expanded with new subcategories and subdivisions. Error tags are listed alphabetically, except for the 'Other' tags, which are excluded in this extension. **The table includes abbreviations:** *Chr. for 'Characters' *Fem for 'Feminine' *Masc. for 'Masculine' *Sing. for 'Singular' *Pl. for 'Plural' *Num. for 'Number' *Conj. for 'Conjunctions' *Nom. for 'Nominative' *Juss. for 'Jussive' *Pron. for 'Pronouns' *Acc. for 'Accusative' *C for 'Confusion' and *E for 'Error(s)'. **Morph:** Morphology, Class: Error Class, Tag: Error Tag, Sub2: Subclass of a subclass, Sub3: Subclass of a subclass of a subclass.

Morphological Template	Description
Template 1: فعل (fa'ala)	Expresses the general verbal meaning of the root in question.
Template 2: فَعَّلَ (fa''al)	Built on template 1 by doubling the middle radical (adding a shadda to it). Often is an intensive or causative version of template 1.
Template 3: فاعل (faa'il)	Built on template 1 by adding an alif between the first and second radicals. It gives an transitive or indicates a relation meaning to the form 1 verb, describes someone doing the act to or with someone else.
Template 4: أفعل (af'al)	Built on template 1 by prefixing an alif and putting a sukuun over the first radical. Similar to template 2 in that it is usually a causative or transitive version of template 1.
Template 5: تفعَّل (taff'al)	Built on template 2 by adding the prefix تـ (ta). Often a reflexive or passive version of template 2.
Template 6: تفاعل (tafaa'al)	Built on template 3 by adding the prefix تـ (ta) . Usually a reflexive or passive version of template 3.
Template 7: انفعال (infa'ala)	Built on template 1 by adding the prefix انـ (inna). Usually a reflexive and/or passive version of template 1.
Template 8: افتعل (ifta'la)	Built on template 1 by adding the prefix اـ (?a) and placing a sukuun must be placed over its first radical.It indicate a reflexive nuances or doing something intentionally version of the template 1.
Template 9: افعلاً (if'all)	Built on template 1 by adding the prefix اـ (?a), placing a sukuun over its first radical, and adding a shadda to the last radical.
Template 10: استفعال (istf'al)	Built on template 1 by adding the prefix استـ (?sta) and inserting a (ta) between the first and second radicals, a sukuun must be placed over the first radical. Often a considerative version of template 1, means to consider or to deem someone in relation to template 1.

Table D.2: Synthetic error generation templates for word inflection based on the ATB.

E ATB Synthetic Data

Orthographic Errors

- **Missing Characters (OM)** In generating OM errors, we apply three distinct types of rules. The initial approach involves randomly removing a character, with each character having an equal probability of being removed, or following a distribution where certain characters like م, ي, ئ, ل are more likely to be removed than others. The second rule focuses on removing a character in instances of repetition. For example: لعملية would be لعلية and جيا would be جيا

The third rule involves removing words that are written but not pronounced, such as: الألف Hamzet al wasel الوصل الشمسية and و همزة الوصل after ف, Alif Tanwin التنونين and ر. and ئ.

- **Wrong Order of Word Characters (OC)** When generating OC errors, we either shuffle two adjacent characters within the same word or find a word that shares the same characters as the target word aimed to change and has a

Levenshtein distance of 2. For example يجب would be جيب

- **Replacement in word characters (OR)** In generating OR errors, we pursue two distinct methods. The first one, we seek a word with a Levenshtein distance of 1 from the target word. For the second one, we randomly select a character and replace it with another character based on specific criteria: 1- Characters that are visually similar in shape, 2- Characters that are located near each other on the keyboard, 3- Characters that have similar sounds or are phonetically close when pronounced. For example, if we change the character ا, we might select ا, ا, or ا as alternatives. For the character ط, possible substitutions include ت, ظ, ص, ض, ك, or د. For instance, the word البرتقال could be altered to البرتقال. The character ق is replaced with غ, as they are similar in shape and located close to each other on the keyboard.
- **Missing Characters (OM)** In generating OM errors, we follow three distinct techniques. The first technique involves adding

characters that are pronounced but should not be written. For instance, in the case of Alif al mad المد ألف and Tanwin التنوين, the Alif is pronounced in the word هذا, but it should not be written. Thus, the resulting error word is هادا. The second one involves randomly adding characters. For example, the word ففي could be altered to في. The third technique requires searching for a word that has a Levenshtein distance of 1 from the target word and is shorter in length. For example, مع معه could be changed to مع.

Syntax Errors

- **Missing Word (XM)** In generating XM errors, we employ the following strategies: 1- remove pronouns such as demonstrative, reflexive pronouns, ...etc. 2- remove conjunctions, prepositions, or inna ان 3- remove bi-gram collocation words. For instance في في الدور would be altered to الدور.
- **Unnecessary Word (XT)** In XT error generation, we employ the following strategies: 1- Add conjunctions and prepositions. 2- Include unnecessary pronouns. For example, in the phrase نلعب نحن, the pronoun نحن is redundant since it is already indicated by the ن in the verb لعب. 3- Add bi-gram collocation words. For instance, ما يطبع could be modified to ما يحدث يطبع.

Morphological Errors

- In Morphological errors, there are two types of error generated: word inflection and verb tense. Starting with word inflection, in the Arabic language most of the words are derived from three literal root which represents a core meaning or concept. There are cases for four, and five literal root, but the most common one is three literal root. By using the root, a variety of nouns, verbs, adjectives, active participles, and passive participles could be derived by adding specific prefixes, suffixes and using special templates. Each template has a basic meaning associated with the general meaning of the root being used. To derive word inflection errors, we transform any noun, verb, or adjective to another one by changing its template while preserving its meaning by checking the gloss of the original

word and generated inflection. We have used the certain templates to derive inflection on words. For example, An example of inflection المشهورة would be المشهورة الشهيرة. Table D.2 shows the employed templates.

- In Arabic verb tense errors, verbs are modified using specific suffixes and prefixes to indicate gender, number, and person. To transform one verb tense to another, these suffixes and prefixes must be adjusted accordingly. For example, to change the past verb وضعنا to its present form, the suffix نا is removed, and the prefix ن is added. Another example is transforming the past verb التقتنا to its present form تتقيان. Here, the suffix تا (indicating a feminine dual third person) is removed, and the suffix ان (indicating dual form) is added, along with the prefix ت to indicate a feminine third person. To convert a verb to future tense, it should first be transformed into the present tense, after which the prefix س is added.

Merge Error In generating merge errors, we consider two main approaches: 1- Randomly merging two consecutive words together. For example, فما كان would be فما كان. 2- Merging specific types of words, which includes: a) Merging prepositions with the following word. For example, عن هذا would be عن هذا. b) Merging ان with the next word. c) Merging negation words with the next word. d) Merging vocative cases with the next word. e) Merging nominal or personal pronouns with the next word. For example, هذه would be هذا الدول.

Split Error In generating split errors, we consider two main approaches: 1- Randomly splitting any word at any position. For example, اهدافه could be split into اهدا فه. 2- Splitting specific types of words, such as: a) Interrogative words: فيما would become ما في. b) Conjunctions: و نضت would be split into نضت و. c) Pronouns: For instance, الفلاحين would be split into الفلاح ين.

Punctuation Error In generating punctuation errors, we consider the following categories: 1- Missing punctuation: We randomly remove punctuation marks from sentences. 2- Unnecessary punctuation: We add random punctuation marks at any position within the sentence. 3- Punctuation confusion: We replace punctuation based on shape and keyboard placement. For example: The comma “,”

is most likely to be replaced with “,” or “;” or “.”. The closing bracket “]” is likely to be replaced with “[”, “(” or “)”.

F Prompt Design

Designing effective prompts is fundamental to leveraging the capabilities of LLMs. A prompt, defined as a set of instructions given to an LLM, acts as a programming tool that tailors the model’s responses to specific tasks (White et al., 2023). By establishing a set of specific rules and intents, prompts can significantly influence both the interactions with the model and its generated outputs. Thus, to achieve the desired outcomes, it is essential to carefully construct prompts that clearly define the task’s objectives and guidelines.

To create effective prompts for our writing tasks, we designed a set of zero-shot prompts tailored to each subtask to elicit the desired responses from the models. For instance, in the I’rab task, each model was instructed to provide the complete I’rab of the sentences and explain their case markers. Similarly, for the MWEs task, models were prompted to identify and explain MWEs within sentences, including their meanings and contextual usage.

We sample 2 sentences for each subtask and use the web interfaces of all the models for the initial evaluation, focusing on the effectiveness of the prompts in eliciting the desired responses. By assessing and comparing the outputs generated by these initial prompts, we identify the most effective prompt. For most subtasks, the prompts we design successfully meet our requirements for the generated outputs.

However, the GEC prompt required a more detailed approach. Initially, we developed a zero-shot prompt instructing the models to identify the type of grammatical error in a given sentence without specifying the error types, using the prompt: "I will give you an MSA sentence with a grammatical error. You should identify the type of error, provide the correct version of the sentence, and explain the error in Arabic. The sentence:". We then use GPT-4o to iteratively refine this prompt, enhancing its ability to identify the type of error from a specified list, provide the correct version of the sentence, and explain the error. After refining, we modify the prompt and standardize it to fit all models. Detailed prompt templates we use can be found in Table G.2.

G Evaluation matrices for Arabic Writing Assistant Tasks

As mentioned earlier how we come up with the evaluation metrics of six tasks of Arabic writing assistance tasks, we provide detailed description of the metrics we employ in our evaluation. For the GEC and explanation tasks for example, we assess three specific criteria: accuracy, clarity, and helpfulness.

Accuracy refers to "the degree to which the response correctly addresses the question or task, without errors in information or execution." This criterion is rated on a scale from 1 to 5 as follows:

- 5: Completely accurate with no errors.
- 4: Mostly accurate with minor errors.
- 3: Moderately accurate with some errors.
- 2: Somewhat accurate with several errors.
- 1: Inaccurate with many errors.

Clarity indicates the ease with which the response can be understood, including how well it conveys the intended message without ambiguity. This criterion is rated on the following scale:

- 5: Extremely clear and easy to understand.
- 4: Mostly clear with minor ambiguities.
- 3: Moderately clear with some ambiguities.
- 2: Somewhat unclear with several ambiguities.
- 1: Very unclear and difficult to understand.

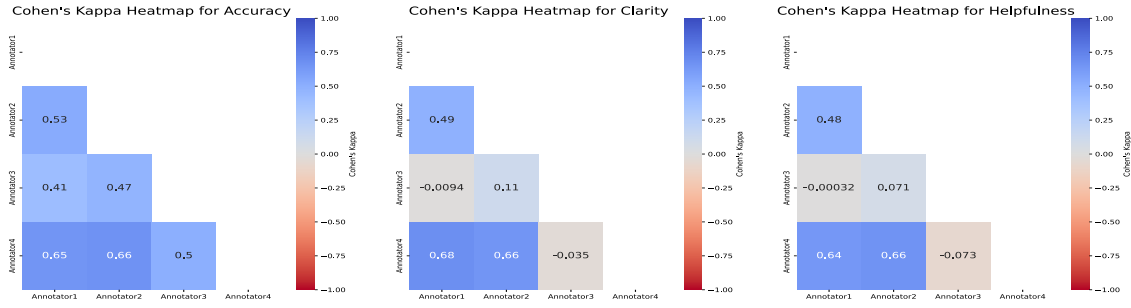
Helpfulness measures "the extent to which the response aids the user in understanding or resolving the issue, providing useful and actionable information." The scale for this criterion is as follows:

- 5: Extremely helpful and informative.
- 4: Mostly helpful with minor gaps in information.
- 3: Moderately helpful with some gaps in information.
- 2: Somewhat helpful with several gaps in information.
- 1: Not helpful and lacking useful information.

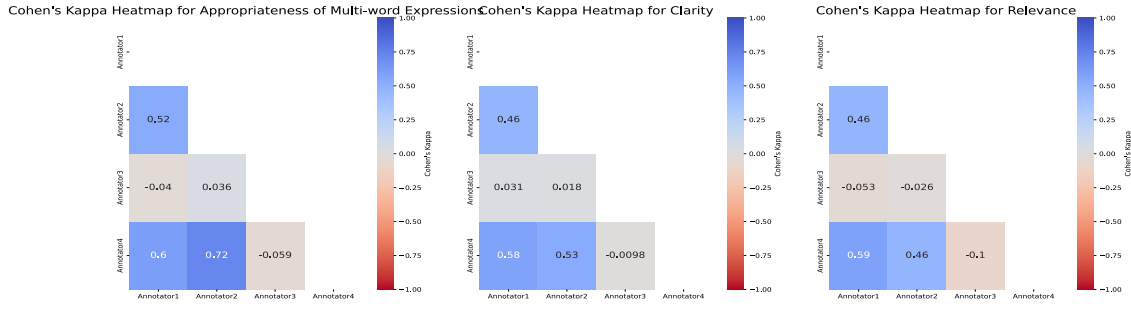
The rest of the metrics are listed in Table G.1. Figure G.1 illustrates the inter-annotator agreement for human evaluation using Cohen's Kappa metrics. Figure G.3 shows examples of responses from the four models across two different tasks.

Gazelle Dataset					
ALC Corpus		New categories		Examples	
Type	Tag	Subclass	Sub2	Text with errors	Corrected text
Orth.	OH	Hamza	Medial Hamza	كانت تلك مسألة هامة.	كانت تلك مسألة هامة.
			Hamzet Wasel	استبشرت بالمعطر خيراً.	استبشرت بالمعطر خيراً.
			The Cutting Hamzah Final Hamza	بخصوص تأمين سكن وطيفي في أحد السكنات لا يجزو المسلم على الكذب.	بخصوص تأمين سكن وطيفي في أحد السكنات لا يجزو المسلم على الكذب.
Syntax	XN	Plural	Masculine Sound Plural Almaqsour Plural	سائقوا السيارات ملتزمون بتعليمات إدارة المرور. وعلى مرمرى طوية زمن منا ونحن لا نعيه، بله لا ندرسه.	سائقو السيارات ملتزمون بتعليمات إدارة المرور. وعلى مرمرى طوية زمن منا ونحن لا نعيه، بله لا ندرسه.
		Number	Composite Numbers Al Aqoud	يمكنكم الآن مشاهدة الحلقة الثالثة عشر حضر مريض في الخمسون من عمره.	يمكنكم الآن مشاهدة الحلقة الثالثة عشرة حضر مريض في الخمسين من عمره.
Semantics	SW	Word Selection Error	Adding Unnecessary Prepositions	تحري فلان عن الأمر.	تحري فلان الأمر.
			Removing Necessary Prepositions	تسعى الدولة تحقيق مكاسب دولية.	تسعى الدولة إلى تحقيق مكاسب دولية.
	SF	Conjunction Error	Replacing Prepositions	ثم وجدت طلبيا في بلاد مختلفة.	ثم وجدت طلبيا من بلاد مختلفة.
			Adding Unnecessary Conjunction Removing Necessary Conjunction	ووصلت صدى هذه الأصوات حتى إلى قلوب بعض المسلمين والمسلمات. اذهب إلى فلان وقل له كذا.	ووصلت صدى هذه الأصوات إلى قلوب بعض المسلمين والمسلمات. اذهب إلى فلان فقل له كذا.
Punc.	PO	Other Punctuation	Comma (.)	فضلاً عن ذلك، غير عن رايه بأسلوب سلس.	فضلاً عن ذلك، غير عن رايه بأسلوب سلس.
	PC	Commission Errors	Full Stop (.)	الشمس تزود الأرض بالحرارة والضوء ؟	الشمس تزود الأرض بالحرارة والضوء.
	PM	Misuse of Punctuation Marks	Question Marks (?)	هل يستطيع هذا القس أن يعلن رغبته	هل يستطيع هذا القس أن يعلن رغبته ؟

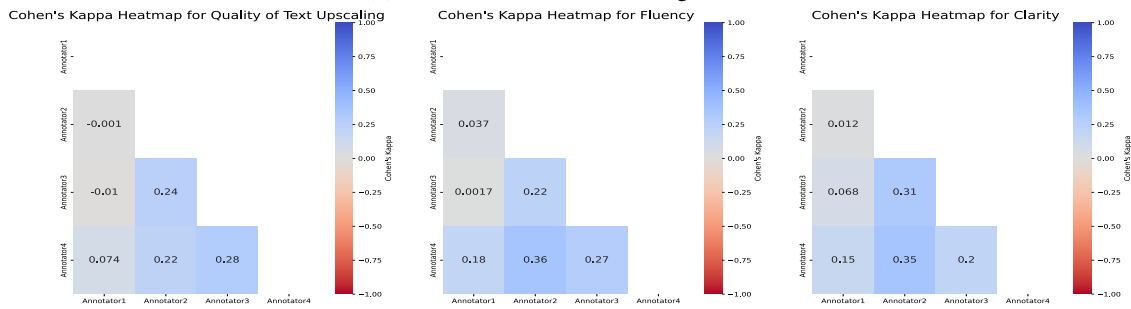
Table F.1: Comparison of *Gazelle* with examples taken from the QALB 2014, QALB 2015, and ZAEBUC datasets, all of which are based on the ALC error taxonomy. *Gazelle* provides more fine-grained and specific error type coverage within this taxonomy. *Sub2 represents our extended subclass within the ALC error types. *Orth.: Orthography *Punc.: Punctuation.



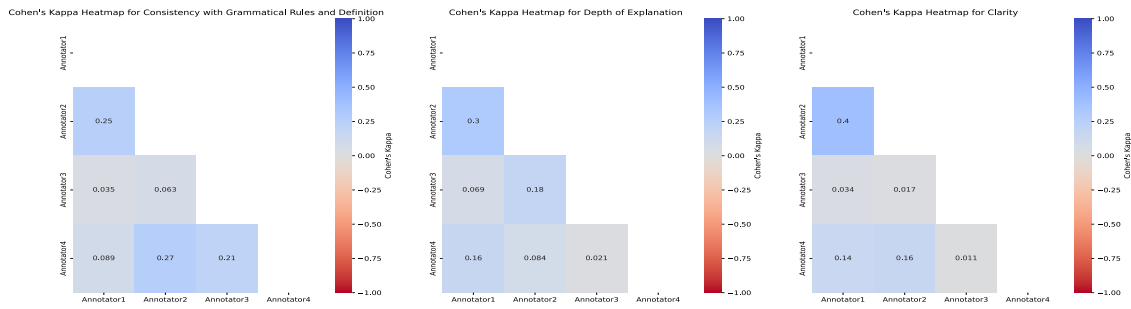
(a) GEC+ Explanation task inter-annotators agreement.



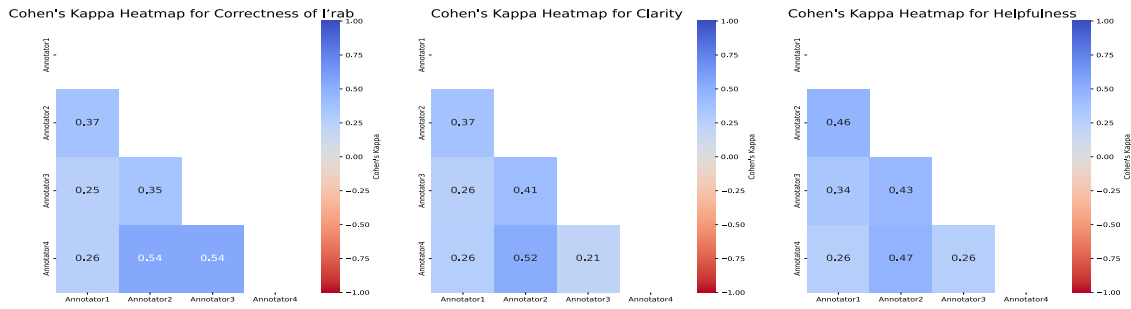
(b) MWE task inter-annotators agreement.



(c) Text refinement task inter-annotators agreement.



(d) Grammatical Rules task inter-annotators agreement.



(e) I'rab task inter-annotators agreement.

Figure G.1: Inter-annotator agreement for human evaluation using Cohen's Kappa.

Grammatical Error Correction & Explanation

Correct any grammatical errors in the given sentence, ensuring it is written correctly in Arabic. Identify and assign the error and type of errors to the provided list of errors, explain the error, and be more specific.

\n

The sentence: الحَقُّ يعلو ولا يُعلا عليه.

\n\n

List of error types:

- Orthographic
- Syntax
- Morphology
- Semantics
- Punctuation
- Merge
- Split

The output should be in the following JSON format:

```
{
  "correction": corrected sentence,
  "error": error in the sentence,
  "error_type": error in the sentence from the list of error types with more details,
  "explanation": explanation of the error.
}
```

Multi Word Expressions and Metaphors

Correct the errors in the sentence and specifically the use of multi-word expressions or idioms. Explain the error and identify the part containing it.

The sentence: "مع تقدم العمر وتراكم المسؤوليات، تسللت منها هموم لا مفر منها، تفرض نفسها على كل جانب من جوانب حياتها."

Grammatical Rules & Definition

Answer the following question about Arabic grammatical rules and definitions in Arabic.

The question: ما هو الفعل الأجوف؟

I'rab

حدد الإعراب الكامل للجملة المعطاة.
الجملة: ما جاء التلاميذ سوى تلميذين

Text- Up Scaling

Rewrite the following text in formal Arabic, being careful to avoid colloquial words.

الجملة: و المدام أيضاً جاءت بخصوص الإعلان الذي في الجريدة؟

Figure G.2: Prompts used for LLMs evaluation.

Task	Measures	Scale Details
Grammar Error Correction+Explanation	Accuracy	5: Completely accurate with no errors. 4: Mostly accurate with minor errors. 3: Moderately accurate with some errors. 2: Somewhat accurate with several errors. 1: Inaccurate with many errors.
	Clarity	5: Extremely clear and easy to understand. 4: Mostly clear with minor ambiguities. 3: Moderately clear with some ambiguities. 2: Somewhat unclear with several ambiguities. 1: Very unclear and difficult to understand.
	Helpfulness	5: Extremely helpful and informative. 4: Mostly helpful with minor gaps in information. 3: Moderately helpful with some gaps in information. 2: Somewhat helpful with several gaps in information. 1: Not helpful and lacking useful information.
Multi-word Expressions	Appropriateness of Multi-word Expressions	5: Extremely appropriate and contextually fitting. 4: Mostly appropriate with minor issues. 3: Moderately appropriate with some issues. 2: Somewhat appropriate with several issues. 1: Inappropriate and contextually unfitting.
	Clarity	5: Extremely clear and easy to understand. 4: Mostly clear with minor ambiguities. 3: Moderately clear with some ambiguities. 2: Somewhat unclear with several ambiguities. 1: Very unclear and difficult to understand.
	Relevance	5: Completely relevant to the task/question. 4: Mostly relevant with minor off-topic elements. 3: Moderately relevant with some off-topic elements. 2: Somewhat relevant with several off-topic elements. 1: Irrelevant to the task/question.
I'rab	Correctness of I'rab	5: Completely correct with no errors. 4: Mostly correct with minor errors. 3: Moderately correct with some errors. 2: Somewhat correct with several errors. 1: Incorrect with many errors.
	Clarity	5: Extremely clear and easy to understand. 4: Mostly clear with minor ambiguities. 3: Moderately clear with some ambiguities. 2: Somewhat unclear with several ambiguities. 1: Very unclear and difficult to understand.
	Helpfulness	5: Extremely helpful and informative. 4: Mostly helpful with minor gaps in information. 3: Moderately helpful with some gaps in information. 2: Somewhat helpful with several gaps in information. 1: Not helpful and lacking useful information.
Grammatical Rule + Definitions	Consistency with Grammatical Rules and Definitions	5: Completely consistent and correct. 4: Mostly consistent with minor inconsistencies. 3: Moderately consistent with some inconsistencies. 2: Somewhat consistent with several inconsistencies. 1: Inconsistent and incorrect.
	Depth of Explanation	5: Extremely thorough and detailed. 4: Mostly thorough with minor details missing. 3: Moderately thorough with some details missing. 2: Somewhat thorough with several details missing. 1: Superficial and lacking in detail.
	Clarity	5: Extremely clear and easy to understand. 4: Mostly clear with minor ambiguities. 3: Moderately clear with some ambiguities. 2: Somewhat unclear with several ambiguities. 1: Very unclear and difficult to understand.
Text Upscaling	Quality of Text Upscaling	5: Extremely high quality with significant improvement. 4: High quality with notable improvement. 3: Moderate quality with some improvement. 2: Low quality with minimal improvement. 1: Poor quality with no improvement or degradation.
	Fluency	5: Extremely fluent and natural. 4: Mostly fluent with minor awkwardness. 3: Moderately fluent with some awkwardness. 2: Somewhat fluent with several awkward phrases. 1: Not fluent and very awkward.
	Clarity	5: Extremely clear and easy to understand. 4: Mostly clear with minor ambiguities. 3: Moderately clear with some ambiguities. 2: Somewhat unclear with several ambiguities. 1: Very unclear and difficult understand.

Table G.1: Evaluation criteria we employ for Arabic writing assistance tasks in our work.

Online Sources

<http://www.al-musheer.com/play-735.html>

<https://www.wuduh1.com/2023/09/constructive-style.html>

<https://learning.aljazeera.net/ar/grammar>

<https://loghate.com/>

<https://www.twinkl.ae/teaching-wiki/hmzt-alwsl-whmzt-alqt>

Table G.2: List of some of the publicly available online sources used in collecting *Gazelle* dataset.

	Prompt	GPT-4o response	Cohere Command R+ response	GPT-4 response	Gemini 1.5 Pro response
<p>GEC + Explanation</p> <p>Correct any grammatical errors in the given sentence, ensuring it is written correctly in Arabic. Identify and assign the error and type of errors to the provided list of errors, explain the error, and be more specific.</p> <p>In</p> <pre>{ "error": "Syntax - Gender Agreement", "error_type": "Syntax - Gender Agreement", "explanation": "The word 'أحد' is feminine, but 'مكونات' is masculine. The correct form should be 'أحد' to match the gender of 'مكونات'." }</pre> <p>The output should be in the following JSON format:</p> <pre>{ "correction": "corrected sentence", "error": "error in the sentence", "error_type": "error in the sentence", "explanation": "explanation of the error." }</pre>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is feminine, but 'مكونات' is masculine. The correct form should be 'أحد' to match the gender of 'مكونات'."</p> <p>... }</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The original sentence is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>
<p>Irab</p> <p>حدد الأخطاء الكاملة للجملة المعطاة "الجملة: أنت نشيط في المدرسة".</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>	<p>أحد مكونات العلية: "العلوية".</p> <p>"error": "أحد".</p> <p>"error_type": "Syntax - Gender Agreement".</p> <p>"explanation": "The word 'أحد' is semantically incorrect because it lacks a subject. The sentence structure suggests that it should have a subject, but it is missing. The corrected sentence adds the subject 'الطالب', which means 'the student' in Arabic, completing the sentence structure and providing a clear and grammatically correct meaning."</p>

Figure G.3: Sample model responses for GEC, explanation, and I'rab tasks.

Task	Accuracy	Clarity	Helpfulness	Appropriateness of MWEs	Relevance	Consistency with Gr. Rules and Definition	Depth of Explanation	Correctness of Frab	Quality of Text Refinement	Fluency
GEC Cohere	2.80	3.28	3.25							
GEC GPT4	2.02	3.17	2.90							
GEC GPT4o	3.27	3.77	3.63							
GEC Gemini15pro	2.70	3.18	3.03							
MWEs Cohere		3.68		3.15	3.78					
MWEs GPT4		3.25		2.63	3.28					
MWEs GPT4o		3.89		3.46	3.91					
MWEs Gemini15pro		3.38		2.91	3.36					
Gr: Rules Cohere		4.20				3.74	4.03			
Gr: Rules GPT4		3.23				2.95	3.11			
Gr: Rules GPT4o		4.73				4.58	4.66			
Gr: Rules Gemini15pro		4.22				4.01	4.33			
Frab Cohere		4.41	4.40					3.91		
Frab GPT4		3.59	3.44					3.01		
Frab GPT4o		4.68	4.71					4.39		
Frab Gemini15pro		3.73	3.83					3.65		
Text Cohere		3.48							3.20	3.39
Text GPT4		3.54							3.15	3.49
Text GPT4o		3.91							3.55	3.70
Text Gemini15pro		3.47							3.39	3.47

Table G.3: Evaluation results for the five subtasks across all LLMs.