Nahw: A Comprehensive Benchmark of Arabic Grammar Understanding, Error Detection, Correction, and Explanation

Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed
Qatar Computing Research Institute, HBKU, Qatar
{hmubarak, mhawasly, abumohamed}@hbku.edu.qa

Abstract

Grammar comprehension is a critical capability for large language models (LLMs) to achieve fluency in a target language. In low-resource settings, such as the case with Arabic, limited availability of high-quality data can lead to significant gaps in grammatical understanding, making systematic evaluation essential. We introduce ***Nahw***, a comprehensive benchmark for Arabic grammar that covers both theoretical knowledge and practical applications, including grammatical error detection, correction, and explanation. We evaluate a range of LLMs on these tasks and find that many models still exhibit substantial deficiencies in Arabic grammar comprehension, with GPT-4o achieving a score of 67% on average over all tasks, while the best performing Arabic model in our experiment (ALLaM-7B) achieving 42%. Our experiments also demonstrate that while fine-tuning with synthetic data can improve performance, it does not match the effectiveness of training on natural, high-quality data.

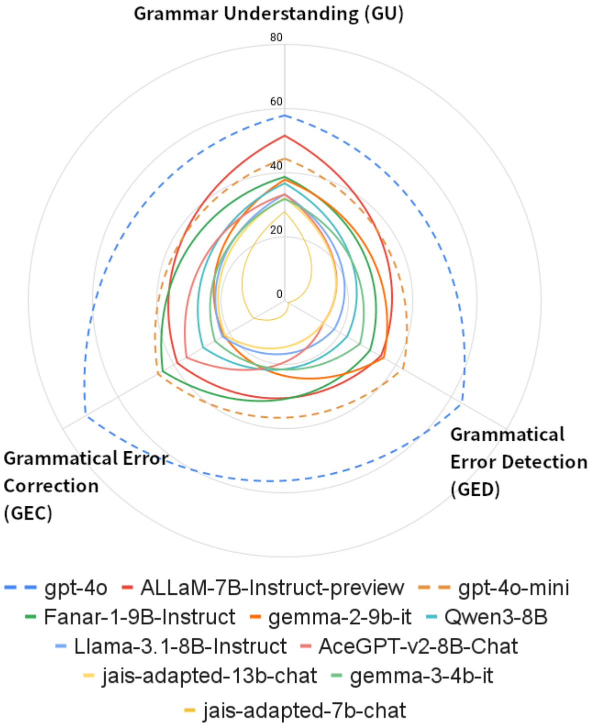


Figure 1: Summary radar plot of the results of the Arabic grammar understanding, error detection and correction tasks on a selection of Arabic and multilingual medium-sized LLMs. Larger LLMs (`GPT-4o` and `GPT-4o-mini`) are shown for reference with dashed lines.

1 Introduction

Mastering language is a fundamental strength of large language models (LLMs), particularly when they are intended for educational use. Language proficiency covers a wide range of skills, with grammar being a core component. Assessing how well LLMs understand grammatical structures is essential – especially for languages that are hindered by limited high-quality training data like Arabic.

Despite growing interest in Arabic NLP, grammar-focused resources remain scarce. Existing corpora often address individual linguistic aspects like spelling or diacritization, but rarely provide explanations of grammatical errors, and educational datasets suitable for evaluating model reasoning about grammar are rare. Developing datasets and benchmarks that capture Arabic's grammatical complexity is therefore crucial to assess, improve, and adapt LLMs for Arabic language education and grammatical feedback applications.

To address this limitation, we present ***Nahw***[1], the first comprehensive resource for benchmarking Modern Standard Arabic (MSA) grammar understanding, error detection, correction and explanation. Our contributions include:

[1] Nahw نَحْو stands for *Arabic grammar*.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6310–6328
March 24-29, 2026 ©2026 Association for Computational Linguistics

1. We introduce a new hierarchy of grammar comprehension tasks, comprising general understanding of theory and application, error detection, error correction and explanation.

2. We introduce ***Nahw*-MCQ**, a standard Arabic grammar MCQ dataset with 5K natural questions, answers, explanations, and metadata.

3. We introduce ***Nahw*-Passage**, a grammatical error correction and explanation dataset of 100 passages in modern standard Arabic, annotated with grammatical and morphological errors, their corrections, in addition to explanations.

4. We benchmark Arabic-centric and multilingual open-weight models, in addition to large proprietary models (namely `GPT-4o` and `GPT-4o-mini`) on the proposed tasks.

5. We show that fine-tuning with synthetic data can enhance small models' grammatical comprehension.

We release the benchmarking data (*Nahw*-MCQ and *Nahw*-Passage), all the prompts, and 10K synthetic MCQ dataset to support future work, on `https://github.com/qcri/nahw-arabic-grammar-benchmark/`.

2 Background on Arabic Grammar

Arabic grammar (النحو العربي, *An-Nahw Al-'Arabi*) defines the structural and morphological principles governing word formation and sentence composition in Standard Arabic. It establishes rules for **case endings** (إعراب, *I'raab*) that mark the syntactic role of words—most prominently the nominative, accusative, and genitive cases (Ryding, 2005). Also, it requires that verbs must agree with their subjects in person, number, and gender, while nouns and adjectives require full agreement in number, gender, definiteness, and case (Watson, 2002).

Arabic further includes a wide range of particles—such as prepositions, adverbs, and complementizers (e.g., "that", "if", ”to”)—that alter the case endings of the words they govern (Haywood and Nahmad, 1965). Nouns occur in three numbers (singular, dual, and plural), and are inflected with distinct case endings. Generally, the case marker reflects the syntactic function of the word in the sentence, encoding grammatical relations that are often implicit in other languages. For example, in this sentence أكلتِ البنتُ التفاحةَ ("ate-*t* the-girl-*u* the-apple-*a*", the girl ate the apple), the verb should be suffixed with a subject feminine singular marker *"t"*, and the subject should have a nominative case ending marker *"u"* to indicate its role in the sentence. Similarly, the object should have an accusative marker *"a"*.

Morphologically, Arabic is a templatic and derivational language where words are generated from consonantal roots using patterned vocalic and affixal structures (Habash, 2010). This root-and-pattern system creates a strong interaction between morphology and syntax, making Arabic grammar particularly rich and complex. The common omission of diacritics or short vowels in modern Arabic text further amplifies ambiguity, posing additional challenges for language learners and computational models alike (Madi and Al-Khalifa, 2018). Consequently, automatic grammar understanding, error detection, and correction in Arabic remain open research problems with substantial linguistic complexity.

3 Related Works

Arabic grammar correction in NLP Research has gained momentum in recent years, paralleling the advances in neural and transformer-based models. Prior to that, the early work relied on rule-based and statistical methods — e.g., rule-driven grammar checkers (Shaalan, 2005), ontology / parser-based and constraint-generation methods (Moukrim et al., 2021), and dependency-grammar models combined with decision-tree classifiers (Alothman and Alsalman, 2020) — which reported high detection rates on vowelized and non-vowelized corpora.

The introduction of shared tasks such as QALB-2014 (Mohit et al., 2014) and QALB-2015 (Rozovskaya et al., 2015) marked a turning point for Arabic error correction research by providing standardized datasets and evaluation benchmarks. These initiatives stimulated the development of a wide range of systems, including rule-based, statistical, and hybrid approaches. However, grammatical and syntactic errors constitute only 3% of the errors in QALB corpus, with the vast majority of annotations addressing orthographic errors, particularly those involving confusions of Hamza (ء), Taa Marbouta (ة) vs. Haa (ه), and Yaa (ي) vs. Alif Maqsoura (ى).

Recent studies have leveraged pre-trained transformer-based language models, such as AraBERT (Antoun et al., 2020), mBERT (Devlin et al., 2019), and AraT5 (Nagoudi et al., 2022), for both Arabic grammatical error detection and grammatical error correction. These models, typically fine-tuned on Arabic learner corpora and synthetically generated error datasets, have achieved state-of-the-art performance, with F1 scores often exceeding 70% on the QALB corpus.

Synthetic data generation, including back-translation and error tagging, has been widely adopted to address data scarcity in Arabic grammar (Alrehili and Alhothali, 2025; Ismail et al., 2025; Abdelrehim et al., 2025), a persistent challenge for morphologically rich and low-resource languages like Arabic. Tools such as ARETA (Belkebir and Habash, 2021) have been developed for automatic error type annotation, further supporting corpus creation and system evaluation.

Instruction-finetuned large language models (LLMs) like `GPT-4` have shown promise in Arabic grammar error correction, especially when combined with few-shot learning and expert prompting (Nagoudi et al., 2023). Data augmentation using LLMs like ChatGPT, sequence-to-sequence transformers, or rule-based systems has enabled the creation of large-scale corpora, such as the Tibyan corpus (Alrehili and Alhothali, 2024), which includes a wide range of error types and supports robust model training.

Overall, the field has progressed from rule-based and statistical methods to advanced neural and transformer-based approaches, with ongoing efforts to address data scarcity, improve error annotation, and enhance system performance for both native and non-native Arabic texts.

4 Grammar Comprehension Task Hierarchy

We introduce a hierarchy of four tasks designed to evaluate different dimensions of grammatical proficiency, with proposed implementations and metrics. These tasks are:

- **Grammar Understanding (GU)**: This task assesses the model's theoretical or applied understanding of grammar. It is formulated as a multiple-choice question (MCQ) task, where the model selects the correct answer from four options. Performance is measured using accuracy.

- **Grammatical Error Detection (GED)**: In this task, the model is given a short text and asked to identify erroneous words. It is implemented as a generation task and evaluated using the F1 score against the gold annotations of the actual errors.

- **Grammatical Error Correction (GEC)**: Here, the model is presented with a short text containing a single highlighted grammatical error and asked to generate the correct form of the erroneous word. This is a generation task evaluated by accuracy.

- **Grammatical Error Explanation (GEX)**: Extending the previous task, the model is given a sentence with a highlighted error and its correction, and is asked to explain the reason for the correction. This generation task is manually evaluated by an expert, and we developed a judging rubric for this task that could be found in Appendix D. We investigate in this paper the feasibility of using LLM-as-a-judge for scoring using the same rubric.

5 Dataset Construction

In this work we focus on Modern Standard Arabic as the only form of Arabic whose grammar is formally taught in Arabic-speaking countries, and whose syntactic and morphological rules are standardized and identical in all curricula.

5.1 Grammar Understanding task

We collected natural Arabic grammar questions from the educational website **https://www.alnahw.com** ("*The Grammar*"), one of the most widely used Arabic educational websites specializing in teaching Arabic grammar to school-level students, after obtaining their formal approval and paying a licensing fee to acquire, use, and share their data for research purposes. The platform hosts more than 15K questions prepared by more than 20 teachers, and supports more than 20K active students at the elementary, preparatory, and secondary levels. In addition to grammar questions, the website provides educational articles, books, and videos in various formats.

For the Arabic grammar understanding task we focus on multiple choice questions (MCQs) as the most structured and pedagogically relevant format for evaluating grammar understanding. The original questions were provided in more than 200 plain

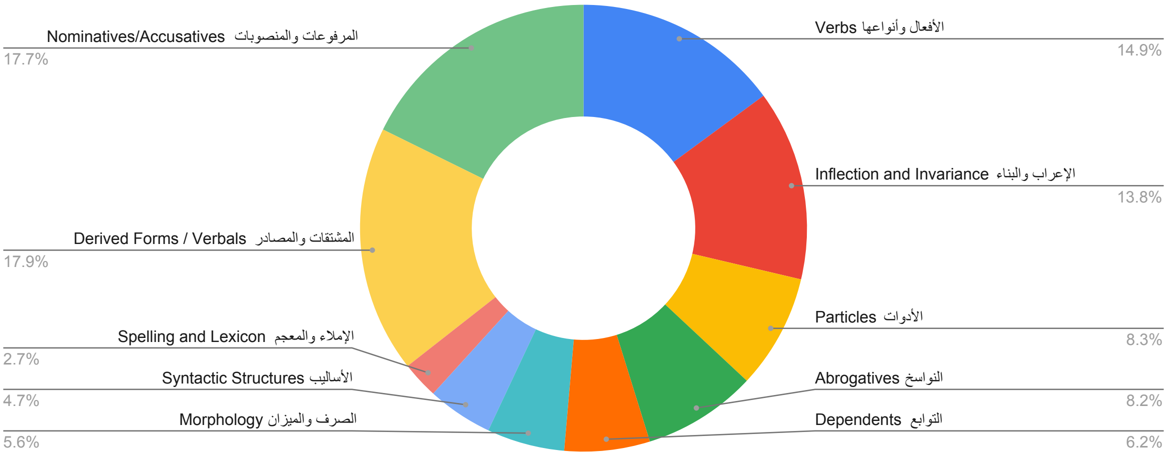


Figure 2: Topic composition of *Nahw*-MCQ dataset

text files (see Figure 4 in Appendix A for an example) that required format standardization (some questions featured fewer or more than four options, broken or incomplete items, or control characters). We wrote a script in Python to normalize the data into a unified structure where each question has four options with only one correct answer in addition to an explanation. After removing duplicates, we extracted a total of 5K MCQs to form the test set *Nahw*-MCQ. To ensure completeness, a trained linguist reviewed the data and verified the final format.

While the content of the questions follow the Egyptian national curriculum – according to the source – the questions themselves are novel: they do not appear verbatim in the original textbooks, but written from scratch by expert professional Arabic grammar teachers inspired by the curriculum, with additional metadata:

- **Grade**: ranging from Grade 6 to Grade 12.
- **Lesson**: a categorization into 80 topics covering the core aspects of Arabic grammar and morphology.
- **Difficulty Level**: Easy, Medium, or Hard.
- **Explanation**: A justification of the correct answer and the reasons for rejection the other distractor options.

Additionally, we prompted `GPT-4o` to classify the questions into either **Practical** or **Theoretical**. An expert linguist evaluated a sample of 200 random responses and reported an accuracy of 96%. The distribution of the questions across the grade, difficulty and type metadata dimensions is presented in Table 1.

Also, we prompted `GPT-4o` to cluster the lessons into *topics*. This resulted in a taxonomy of 10 major topics that was later reviewed by an expert linguist. The prompt and the resulting taxonomy are in Appendix B. We use this coarser categorization later to report aggregate performance in the grammar understanding task (GU). Topic distribution in the dataset can be found in Figure 2.

	Attribute	#Quest.	%
Grades	Grade 6	294	6
	Grade 7	157	3
	Grade 8	307	6
	Grade 9	350	7
	Grade 10	354	7
	Grade 11	501	10
	Grade 12	272	5
	Revision (for grades 6-11)	2,765	55
Difficulty	Easy	1,284	26
	Medium	1,685	34
	Hard	2,031	41
Type	Practical Questions	3,040	61
	Theoretical Questions	1,960	39

Table 1: Distribution of *Nahw*-MCQ dataset

5.2 Grammatical Error Detection, Correction and Explanation tasks

In addition to the MCQ dataset, we collected 100 short, free-form Arabic passages from a book on the same website[2], each containing approximately five grammatical or morphological errors with their

[2] Book title: احترف التدقيق اللغوي (*Become a Professional in Proofreading*).

corrections and concise linguistic explanation. We developed a Python script to extract and structure the raw data (See Figure 5 in the appendix for an example). A senior linguist further verified the formatting, completed missing entries, validated the corrections, and cross-checked them against the original source material. This test set (*Nahw-*Passage) has a total of 4,771 words with a total of 511 errors (Avg 5.11 per passage).

6 Results

6.1 Benchmarking

We benchmark a number of LLM models on the four tasks of grammar understanding (GU), error detection (GED), correction (GEC) and explanation (GEX). We chose the following medium-sized Arabic-centric or multilingual open-weight LLMs for benchmarking:

- **`jais-adapted-7b-chat`** and **`jais-adapted-13b-chat`** (Sengupta et al., 2023; Inception, 2024): two instruction-tuned bilingual (Arabic-English) models from Inception AI, built on top of Llama 2.

- **`AceGPT-v2-8B-Chat`** (Liang et al., 2024): a fine-tuned model built on top of Llama 3-8B for Arabic.

- **`ALLaM-7B-Instruct-preview`** (Bari et al., 2025): A high-performance 7B bilingual (Arabic-English) model from SDAIA/Humain AI.

- **`Fanar-1-9B-Instruct`** (Fanar Team et al., 2025): a bilingual (Arabic-English) model from QCRI, built on top of `Gemma-2-9b`.

- **`Gemma-2-9b-it`** (Gemma Team, 2024): an open-weight multilingual model from Google.

- **`Llama3.1-8B-Instruct`** (Grattafiori et al., 2024): the smaller member of the 3.1 series from Meta multilingual models.

- **`Qwen3-8B`** (Qwen Team, 2025): A powerful 8B parameter model from Alibaba Cloud's Qwen3 family.

In addition, we also include results of `GPT-4o` and `GPT-4o-mini` for reference. Figure 1 shows a summary radar plot of the performance of these LLMs in the grammar understanding, error detection and correction tasks. We provide detailed analysis of these results, in addition to grammatical error explanation task, in the next sections.

6.1.1 Grammar Understanding (GU) results

This MCQ task was implemented using LM-Evaluation-Harness (Gao et al., 2024) with the instruction shown in Appendix C.1. Results are shown in Table 2.

As the results show, the highest performance among medium-sized models was achieved by `ALLaM-7B-Instruct-preview`, with an accuracy of 51.38%. While this surpasses the performance of `GPT-4o-mini`, the low rate highlights the complexity of the task and the considerable room for improvement. Hypothetically, if 40% were considered a passing score, only three other models would have passed the Grade 6 test.

Another notable observation is that scores of theoretical questions are consistently higher than practical questions for almost all models. This may reflect the scarcity of high-quality, error-free Arabic content available for training, which limits the models' exposure to practical grammatical usage and fluency.

6.1.2 Grammatical Error Detection (GED) results

The prompt used for this generative task can be seen in Appendix C.2. Post-processing of model outputs was necessary for some models that did not fully follow the prompt formatting instructions. Despite careful handling, some noise in the reported scores is to be expected. The F1 results are presented in the middle column in Table 3.

The best performing model in this task was the multilingual `Gemma-2-9b-it` followed closely by `Allam` then `Fanar`. None of the medium-sized models managed to beat the smaller `GPT-4o-mini` in this task, possibly partly due to superior instruction following in the latter. The generally low scores of medium-sized models on this task suggest deficiencies in precision (returning non-erroneous words), recall (failing to detect some errors), or both, highlighting a critical gap in Arabic grammar comprehension for these models. Moreover, even `GPT-4o` at 64% remains considerably below the performance level of a human linguist.

	Practical	Theoretical	G6	G7	G8	G9	G10	G11	G12	Overall
ALLaM-7B-Instruct-preview	**49.47**	**54.34**	**64.97**	**45.22**	**44.30**	**47.14**	**44.63**	**38.72**	**41.54**	**51.38**
Fanar-1-9B-Instruct	37.73	39.80	45.92	34.39	32.57	41.71	36.16	31.14	34.19	38.54
Gemma-2-9b-it	37.43	38.47	44.22	36.31	37.13	36.86	35.31	33.13	35.29	37.84
Qwen3-8B	36.22	37.55	43.20	36.31	30.62	36.57	32.77	33.33	36.76	36.74
Llama-3.1-8B-Instruct	33.55	32.96	33.67	25.48	33.55	32.86	30.23	33.33	31.99	33.32
AceGPT-v2-8B-Chat	31.32	35.77	35.37	26.11	30.62	36.29	31.07	29.34	29.04	33.06
Jais-adapted-13b-chat	30.92	33.78	35.03	22.29	28.99	31.71	27.12	33.73	29.78	32.04
Jais-adapted-7b-chat	27.99	27.14	28.91	22.93	30.62	25.43	25.99	30.94	27.94	27.66
GPT-4o	52.99	65.61	74.15	56.05	54.72	55.71	45.76	49.50	51.47	57.94
GPT-4o-mini	41.58	48.67	54.42	45.22	39.74	47.14	36.72	38.12	37.87	44.36

Table 2: Accuracy of the grammar understanding task (GU) for a selection of medium-sized LLMs. Larger models (shaded) are added for reference. **Bold** and underlined show the best and the second best results, respectively. G6-G12 are Grades 6 to 12.

	GED	GEC
	F1%	Acc.%
Fanar-1-9B-Instruct	30.78	**43.84**
ALLaM-7B-Instruct-preview	34.62	38.75
AceGPT-v2-8B-Chat	13.72	35.23
Qwen-3-8b	22.47	29.35
Llama-3.1-8B-Instruct	17.90	22.31
Gemma-2-9b-it	**35.74**	21.72
Jais-adapted-13b-chat	13.87	21.53
Jais-adapted-7b-chat	1.31	11.15
GPT-4o	63.90	71.62
GPT-4o-mini	42.61	45.60

Table 3: F1 scores for the grammatical error detection task (GED) and accuracy for the grammatical error correction task (GEC) for a selection of LLMs. **Bold** and underlined show the best and the second best results, respectively.

6.1.3 Grammatical Error Correction (GEC) results

The prompt used for this task is shown in Appendix C.3. The results are in the right-most column of Table 3. Like in GED, post-processing was necessary for this task too, not only due to issues in instruction following by some of the models, but also as a result of variations of the returned corrections. Thus, the noise should be expected in these results too.

For this task, `Fanar-1-9B-Instruct` comes on top with a performance not far from that of `GPT-4o-mini`. On the other hand, `GPT-4o` performs 63% better than the best medium-sized model, but still makes a lot of erroneous or incomplete corrections, especially when the actual correction requires a clarifying diacritic to distinguish it from the mistaken word which the model sometimes ignores.

6.1.4 Grammatical Error Explanation (GEX) results

We benchmarked our best performing medium-sized models, `ALLaM-7B-Instruct-preview` and `Fanar-1-9B-Instruct`, along with `GPT-4o`, on two GEX subtasks using the two parts of the dataset. For *Nahw*-MCQ dataset, each data item includes a question, four options, and justification of the correct choice. Thus, models were asked to explain the correct answer in the MCQ dataset. In *Nahw*-Passage dataset, we present a passage with a single mistake in it along with its correction, and ask the model to explain the correction. Prompts used in benchmarking can be found in Appendix C.4.

Two expert linguists manually and separately evaluated a sampled set of 200 outputs from each model, assigning a score (0–10) based on alignment with reference explanations, with a scoring rubric that can be found in Appendix D. We report the annotation results for the three models in Table 4, and we also report the annotators' Pearson Correlation Coefficient (0.5 to 1.0 = strong positive correlation) of the two annotators to assert the reliability of the evaluation.

As shown in Table 4, `GPT-4o` achieved the highest scores according to the two annotators (8.9/8.8 on MCQs, 6.1/6.8 on passages). Errors produced by `GPT-4o` in GEX task mainly stem from hallucinated or irrelevant explanations, partially correct but incomplete reasoning, or occasional factual inaccuracies.

On the other hand, smaller models (`Allam-7B` and `Fanar-9B`) showed weaker performance (7.1/6.1 and 4.7/4.2 on the MCQ subtask, and

1.7/2.4 and 0.7/0.98 on the passages subtask, respectively). These results indicate that grammatical error explanation in free-text contexts remains highly challenging for current Arabic LLMs.

Model	MCQ	Passage
	(L1, L2), ***Corr***	**(L1, L2),** ***Corr***
GPT-4o	(8.9, 8.8), *0.72*	(6.1, 6.8), *0.66*
ALLaM-7B	(7.1, 6.1), *0.8*	(1.7, 2.4), *0.87*
Fanar-9B	(4.7, 4.2), *0.68*	(0.7, 0.98), *0.71*

Table 4: Human evaluation by two linguists (L1 and L2) of GEX (score out of 10) on *Nahw*-MCQ and *Nahw*-Passage datasets, in addition to annotators' Pearson Correlation Coefficient, for GPT-4o, ALLaM-7B-Instruct-preview, and Fanar-1-9B-Instruct.

LLM-as-a-judge scoring To assess whether GPT-4o can serve as an automatic evaluator for free-text grammatical error explanations, we prompted it to score Allam-7B explanations using the same rubric and the same set of 200 response of the *Nahw*-Passage subtask. The results revealed a substantial discrepancy: while the human expert evaluator assigned an average score of 2.05 to Allam's output, GPT-4o rated the same outputs 4.8 for the same rubric. This gap suggests that GPT-4o's moderate performance in grammatical explanation (average score of 6.45) limits its reliability as an autonomous judge for this task. This highlights the need for further research into developing evaluation models capable of human-level grammatical judgment and critique in order to scale up the GEX task.

6.2 Improving Models in Grammar

To investigate methods for enhancing model performance on the GU task beyond the constraints of scarce natural data, we generated a synthetic MCQ data using GPT-4o for fine-tuning purposes. Subsequently, we fine-tuned Gemma-3-4b-it (Gemma Team, 2025) on this synthetic dataset to assess whether synthetic data generation is a practical approach for improving model comprehension of Arabic grammar. We selected Gemma-3-4b-it because its relatively small size strikes a balance between strong baseline performance and efficient fine-tuning.

6.2.1 Synthetic Data Generation

We generated a collection of 10K synthetic MCQs utilizing GPT-4o. For each lesson included in the benchmark, we constructed a few-shot prompt containing example questions pertaining to that specific lesson and prompted GPT-4o to generate additional questions following the same format. The prompt template employed for synthetic data generation is provided in Appendix E.1.

A language expert evaluated the quality of the synthetic questions and their associated metadata on a sample of 200 records, and assigned an average score of 8.1/10, indicating that GPT-4o can generate high-quality and pedagogically sound synthetic data. Error types in the generated MCQs are listed in Table 10 in the appendix. To support future research and benchmarking efforts, we release this synthetic dataset publicly.

6.2.2 Fine-tuning Results

The results of fine-tuning Gemma-3-4b-it model on the 10K synthetic MCQ dataset are displayed in Table 5, and the training setup and hyperparameters used are listed in Appendix F. Through fine-tuning for 5 epochs, we achieved a 15.8% relative improvement compared to the base model, demonstrating that synthetic data can effectively expose the model to relevant grammatical patterns and question types. We investigated the effect of different sizes of synthetic training data on model performance. The results in Table 5 show a general trend of increasing performance on the GU task as the size of the synthetic training set increases.

Fine-tuning on a smaller collection of 5K new natural examples extracted from the same website resulted in superior performance overall (44.8% relative improvement), suggesting that real, expert-written questions contain richer linguistic signals from which current models derive stronger benefits. These results demonstrate that synthetic data can be utilized to bootstrap model performance, but high-quality natural data remains essential for attaining peak accuracy. Nevertheless, in low-resource scenarios, synthetic data can effectively bridge gaps in model knowledge.

Figure 3 presents the topic-wise accuracy on the GU task for the base Gemma-3-4b-it model and its fine-tuned variants on 10K synthetic and 5K natural examples. Across all major grammar topics, fine-tuning leads to consistent accuracy gains, with the natural data–fine-tuned model achieving the highest performance overall. Notably, the synthetic-data-fine-tuned model shows clear improvements over the base model in every topic, confirming that synthetic augmentation helps the model general-

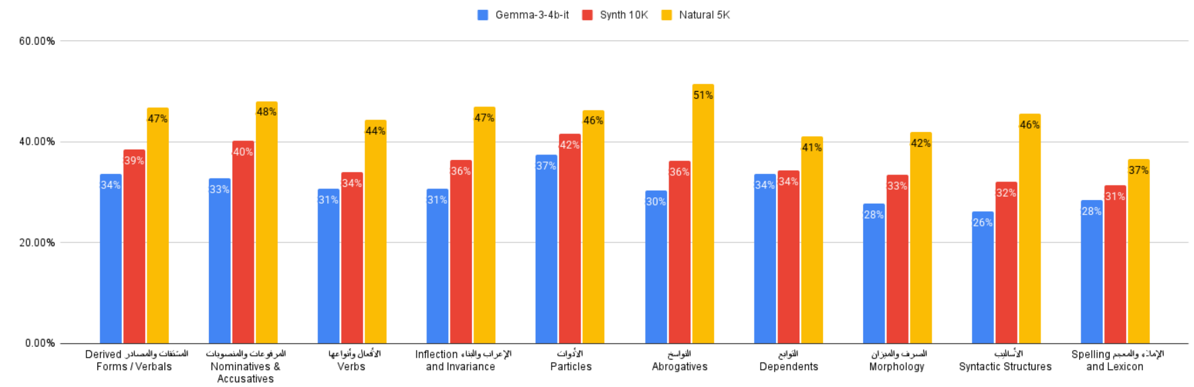


Figure 3: Accuracy at the grammar understanding (GU) task per topic for the base and fine-tuned models. The topics are ordered by their percentage in the benchmark dataset.

	Acc. (%)
`Gemma-3-4b-it`	31.84
`+ Synth. 3k`	32.64
`+ Synth. 7K`	33.74
`+ Synth. 10K`	<u>36.86</u>
`+ Natural 5K`	**46.10**

Table 5: Accuracy of the grammar understanding task (GU) for base and fine-tuned variants of the `Gemma-3-4b-it` model. **Bold** and <u>underlined</u> show the best and the second best results, respectively.

ize to a broad range of grammatical phenomena. The largest gains are observed in high-frequency topics such as "Nominatives & Accusatives," "Particles," and "Inflection and Invariance," which are central to Arabic syntax and morphology. These results highlight that synthetic data can significantly strengthen model performance in a targeted and scalable way, while natural data remains more effective when available.

Model Collapse When finetuning a model to improve its performance on a specific task, it is possible for the model to lose some of its previous capabilities. To ensure that our finetuning did not cause this, we benchmarked the original and the finetuned models on ArabicMMLU[3] (Koto et al., 2024) and the Arabic portion of MMMLU, the professional translation of MMLU (Hendrycks et al., 2021) provided by OpenAI on huggingface[4]. Both of these benchmarks are multi-task and multi-domain MCQ benchmarks that evaluate a model's knowledge and understanding across a wide range of academic subjects, like Arabic linguistics, Math, Islamic studies, history and STEM. The benchmark results are reported in Table 6. The results show that the fine-tuned models sometimes outperform the base model on both benchmarks. This could possibly be attributed to the models gaining a better understanding of Arabic grammar and morphology, thereby improving their ability to comprehend and answer questions more accurately, with more data leading to better performance.

	ArabicMMLU Acc. (%)	**MMMLU/ar** Acc. (%)
`Gemma-3-4b-it`	<u>53.01</u>	39.29
`+ Synth. 10K`	**54.51**	**44.94**
`+ Natural 5K`	51.95	<u>41.16</u>

Table 6: Accuracy in ArabicMMLU and MMMLU/ar for base and fine-tuned variants of the `Gemma-3-4b-it` model. **Bold** and <u>underlined</u> show the best and the second best results, respectively.

7 Conclusion

We introduced ***Nahw***, the first comprehensive benchmark for Arabic grammar understanding, error detection, correction, and explanation. Our contributions include a new hierarchy of grammar tasks that systematically assess different dimensions of grammatical proficiency, the release of *Nahw*-MCQ, a 5K-question dataset covering core grammar concepts, and *Nahw*-Passage, a linguistically annotated error correction and explanation dataset. Benchmarking Arabic-centric and multilingual open-weight and proprietary instruction-tuned LLMs shows that while `GPT-4o` performs best among current models, it still falls short of human expert performance. Medium-sized models

[3] `https://hf.co/datasets/MBZUAI/ArabicMMLU`

[4] `https://hf.co/datasets/openai/MMMLU`

are substantially inferior, revealing major gaps in Arabic grammar competence.

To address this, we proposed a synthetic data augmentation strategy using GPT-4o to generate high-quality grammar questions. Fine-tuning Gemma-3-4b-it on a 10K synthetic dataset yielded a 15.8% relative improvement, while 5K natural examples led to a 44.8% relative gain. Although natural data remains more impactful, these results show that synthetic data can provide substantial and scalable improvements, particularly when natural data is scarce. A language expert rated the synthetic data 8.1/10 on average, confirming its pedagogical soundness.

Future directions include combining natural and synthetic data for fine-tuning, benchmarking fine-tuned models on GED, GEC, and GEX tasks, expanding coverage of grammatical errors from language learners with more complex grammatical phenomena, and developing robust automatic evaluation metrics for grammatical explanation, as even the best current models do not match expert-level reasoning.

Beyond this, while the current benchmark focused on sentence-level grammatical competence, maintaining diagnostic clarity and annotation feasibility, additional skills that involve broader generative or discourse abilities that extend beyond controlled grammar evaluation – such as production and discourse-level consistency – are natural extensions to *Nahw*.

8 Acknowledgment

We would like to thank the owners of **Al-Nahw website** (https://www.alnahw.com) for granting access to their educational resources, their willingness to license the data for research, and their strong commitment to advancing Arabic grammatical error detection, correction, and explanation.

9 Limitations

While this work makes several contributions toward advancing Arabic grammar understanding in LLMs, it has limitations.

First, the dataset was derived from a single source covering only the Egyptian curriculum. Although Arabic grammar rules are shared across Arab countries, lesson focus and difficulty vary by country and teacher. Also, *Nahw*-MCQ and *Nahw*-Passage datasets are relatively small, which may introduce bias.

Secondly, the evaluation focuses on multiple-choice and short free-text tasks, which, while well-structured and linguistically rich, capture only a subset of the ways grammatical knowledge is used in real-world language understanding and generation. Extending evaluation to more open-ended contexts such as essay writing, conversational dialogue, or long-form error correction would provide a more comprehensive assessment of model capabilities.

Thirdly, while we used optimized prompts for benchmarking, different prompt formulations could yield varying results. And while extra care was taken in post-processing the model outputs, noise is still expected in the reported scores.

Fourthly, the synthetic data generation process relied on GPT-4o, which, despite producing high-quality questions, may introduce subtle biases or stylistic regularities that differ from authentic educational material. These artifacts could influence model learning in ways that do not fully generalize to naturally occurring data.

Also, the fine-tuning experiments were conducted on a single small model (Gemma-3-4b-it). While this model is representative of lightweight instruction-tuned architectures, the effectiveness of the proposed synthetic augmentation strategy may differ for larger or smaller models.

Finally, while we conducted human evaluation for explanation quality, automatic evaluation metrics for this task remain underdeveloped. The lack of robust automatic scoring methods limits the scalability of evaluation for free-text grammatical explanations. In addition, due to the cost of human evaluation, annotations of each task were conducted by a single or two linguists on sampled responses; expanding the sample size and annotator pool could influence the findings. The linguists participated in this work were all males from Egypt with 20-30 years of experience in Arabic linguistics.

10 Ethical Considerations

This work involves the collection and release of Arabic grammar datasets and the benchmarking of language models. We took several measures to ensure that the data and methodology respect ethical guidelines.

Data licensing and consent. All natural multiple-choice questions were obtained from alnahw.com under a formal licensing agreement, with the explicit consent of the data owners to use and share the material for research purposes. The passages used for error detection and correction were similarly licensed and reviewed by expert linguists to ensure accuracy and proper attribution.

Data quality and linguistic integrity. To minimize the risk of propagating errors or misleading linguistic patterns, all natural data underwent cleaning, normalization, and expert validation. Synthetic data generated using GPT-4o was manually evaluated by a linguist to assess pedagogical quality and identify systematic errors, which are documented transparently in the Appendix. The expert linguists were compensated fairly by rates verified against regional wage benchmarks from platforms such as Bayt.com and Glassdoor.

Cultural and linguistic sensitivity. Arabic grammar is deeply tied to educational and cultural contexts. Care was taken to focus exclusively on linguistic phenomena and avoid sensitive sociopolitical or religious content. The dataset does not include personally identifiable information (PII) or user-generated content.

Responsible release. We will release the datasets, prompts, and evaluation scripts under a research license to support reproducibility and further work, while discouraging misuse. Model fine-tuning experiments are reported transparently to avoid overstating capabilities or promoting misleading claims about language understanding.

Potential misuse in educational contexts. Because Arabic grammar is a core component of formal education, there is a risk that models fine-tuned or evaluated on this benchmark could be deployed in high-stakes settings, such as automated grading or tutoring, without appropriate human oversight. This could lead to inaccurate assessments, overreliance on imperfect systems, or the reinforcement of educational inequities.

Culture Biases and Language Variety. Although all questions were taken from a single source, this does not introduce grammatical bias. The grammar of Modern Standard Arabic (MSA) is fully standardized across Arabic-speaking countries, and the grammatical rules targeted in our dataset—such as syntax, morphology, and case—are consistent in national curricula everywhere. Differences between countries are mostly in the phrasing of examples, not in the grammar itself. Thus, the dataset represents general MSA grammar, not Egypt-specific Arabic. That said, there might be slight differences between curricula in the assignment of lessons to grades or in the level of complexity that is targeted per grade, but not in the underlying grammatical structures. Although dialects constitute the primary medium of everyday spoken communication across Arabic-speaking communities, they are not fully standardized and often exhibit significant regional variation in grammar, pronunciation, and vocabulary. In contrast, MSA is the only formalized and standardized variety of the language. Our work specifically targets MSA grammar.

Scope. Our goal in this work is to evaluate core grammatical competence, and sentence-level MCQs and short passages are the standard format used in Arabic grammar assessment across educational systems. While evaluating long-form generative abilities is valuable, it adds to assessing grammatical fluency an additional skill dimension that we do not target in this work (discourse planning, coherence, topical development). Our benchmark is designed to isolate grammar-specific errors without confounds from broader generation abilities.

Use of AI in writing. AI tools were used to assist with phrasing and writing refinement throughout the paper, but were not used to generate original research content or new material from scratch.

References

Mohamed Abdelrehim, Marwan Torki, and Nagwa El-Makky. 2025. Hybrid llm and rule-based synthetic data generation for arabic grammatical error correction. In *2025 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, pages 280–285.

Ameerah Alothman and AbdulMalik Alsalman. 2020. An arabic grammar auditor based on dependency grammar. *Advances in Human-Computer Interaction*, 2020(1):8856843.

Ahlam Alrehili and Areej Alhothali. 2024. Tibyan corpus: Balanced and comprehensive error coverage corpus using chatgpt for arabic grammatical error correction. *Preprint*, arXiv:2411.04588.

Ahlam Alrehili and Areej Alhothali. 2025. Towards the development of balanced synthetic data for correcting grammatical errors in arabic: An approach based on error tagging model and synthetic data generating model. *Preprint*, arXiv:2502.05312.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Gemma Team. 2024. Gemma.

Gemma Team. 2025. Gemma 3.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

JA Haywood and HM Nahmad. 1965. *Arabic grammar of the written language*. Lund Humphries.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Inception. 2024. Jais family model card.

Karim Ismail, Sherif Abdou, Mohamed Farouk, and Ahmed Salem. 2025. Transformers to the rescue: alleviating data scarcity in arabic grammatical error correction with pre-trained models. *Neural Computing and Applications*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Nora Madi and Hend S Al-Khalifa. 2018. Grammatical error checking systems: A review of approaches and emerging directions. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 142–147. IEEE.

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

Chouaib Moukrim, Tragha Abderrahim, Almalki Tarik, et al. 2021. An innovative approach to autocorrecting grammatical errors in arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 33(4):476–488.

El Moatez Billah Nagoudi et al. 2022. Arat5: A sequence-to-sequence model for arabic text correction. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*.

El Moatez Billah Nagoudi et al. 2023. Arabicgpt and arallm: Evaluating instruction-tuned large language models on arabic nlp tasks. In *Proceedings of the Eighth Arabic Natural Language Processing Workshop (WANLP)*.

Qwen Team. 2025. Qwen3.

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

Karin C Ryding. 2005. *A reference grammar of modern standard Arabic*. Cambridge university press.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

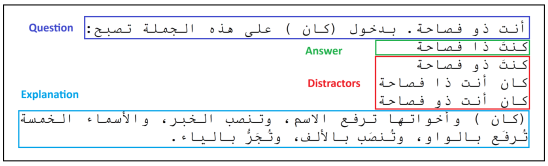
Khaled F Shaalan. 2005. Arabic gramcheck: A grammar checker for arabic. *Software: Practice and Experience*, 35(7):643–665.

Janet CE Watson. 2002. *The phonology and morphology of Arabic*. OUP Oxford.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Example Raw Data

Figures 4 and 5 show examples of the acquired raw data from which *Nahw* was constructed.



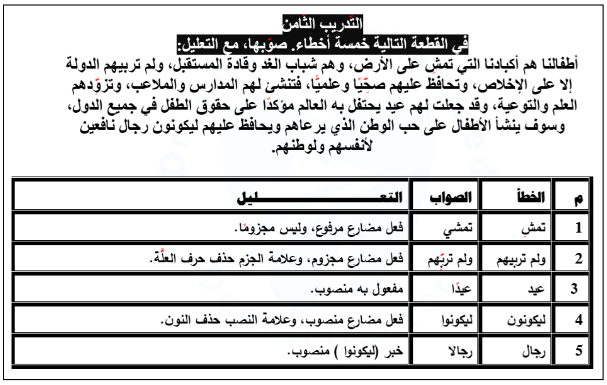
Question أنت ذو فصاحة. بدخول (كان) على هذه الجملة تصبح:

Answer كنتَ ذا فصاحة

Distractors
كنتَ ذو فصاحة
كان أنت ذا فصاحة
كان أنت ذو فصاحة

Explanation
(كان) وأخواتها ترفع الاسم، وتنصب الخبر، والأسماء الخمسة
تُرفَع بالواو، وتُنصَب بالألف، وتُجَرُّ بالياء.

Figure 4: An example of the raw MCQ data which shows the question, correct answer, distractors, and an explanation. This is an example of a *Hard* question from *Grade 6* that covers the topic of كان وأخواتها (the incomplete verb Kana and sisters).
Translation: You are eloquent. By adding (the incomplete verb “Kana”) to the sentence, it becomes: You were eloquent/... (options with different grammar case marker for nominative and accusative cases), followed by an explanation about the rule of applying incomplete verbs; nominative subject and accusative predicate.



التّدريب الثّامن

في القطعة التالية خمسة أخطاء. صوّبها، مع التعليل:

أطفالنا هم أكبادنا التي تمشِ على الأرض، وهم شباب الغد وقادة المستقبل، ولم تربيهم الدولة إلا على الإخلاص، وتحافظ عليهم صحّيًا وعلميًّا، فتنشئ لهم المدارس والملاعب، وتزوّدهم العلم والتوعية، وقد جعلت لهم عيد يحتفل به العالم مؤكدًا على حقوق الطفل في جميع الدول، وسوف ينشأ الأطفال على حب الوطن الذي يرعاهم ويحافظ عليهم ليكونون رجال نافعين لأنفسهم ولوطنهم.

م	الخطأ	الصواب	التعـــــــــــــليل
1	تمشِ	تمشي	فعل مضارع مرفوع، وليس مجزومًا.
2	ولم تربيهم	ولم تربِّهم	فعل مضارع مجزوم، وعلامة الجزم حذف حرف العلّة.
3	عيد	عيدًا	مفعول به منصوب.
4	ليكونون	ليكونوا	فعل مضارع منصوب، وعلامة النصب حذف النون.
5	رجال	رجالا	خبر (ليكونوا) منصوب.

Figure 5: An example of the raw passages, showing the passage text, a table of extracted errors, corrections, and explanations.

Translation: Our children are our souls that **walk** on the ground, and they are the youth of tomorrow... (the bold verb is written in Arabic in subjunctive form while it should be in nominative form (indicative mood)).

B Lesson Taxonomy

B.1 Prompt

The prompt used to generate the taxonomy is shown below.

> **Taxonomy prompt**
>
> I'll give you a list of Arabic grammar lesson titles. I need you to organize and categorize these classes into a taxonomy. Make sure to include all the lessons. Make sure that your taxonomy includes all the items I mentioned and that it does not change the naming of any class in any way.
>
> *<lesson titles>*

B.2 Generated taxonomy

Tables 7 and 8 show the suggested taxonomy and lesson classification created by GPT-4o and corrected and approved by an expert linguist.

Table 7: Taxonomy of grammar lessons

الأدوات	أدوات النفي وأدوات الاستفهام
الأدوات	أسماء الإشارة والضمائر المنفصلة
الأدوات	أنواع ما
الأدوات	أنواع من
الأدوات	الأدوات
الأدوات	الضمائر المنفصلة والمتصلة
الأدوات	الظرف
الأدوات	كم الاستفهامية والخبرية
الأدوات	لا النافية للجنس
الأساليب	أسلوب الاختصاص
الأساليب	أسلوب التعجب
الأساليب	أسلوب الشرط
الأساليب	أسلوبا الإغراء والتحذير
الأساليب	أسلوبا المدح والذم
الأفعال وأنواعها	أسماء الأفعال
الأفعال وأنواعها	إسناد الأفعال إلى الضمائر
الأفعال وأنواعها	الأفعال الخمسة
الأفعال وأنواعها	الصحيح والمعتل من الأفعال
الأفعال وأنواعها	الفعل التام، والفعل الناقص
الأفعال وأنواعها	المجرد والمزيد
الأفعال وأنواعها	بناء الفعل للمجهول
الأفعال وأنواعها	توكيد الأفعال
الأفعال وأنواعها	جزم الفعل المضارع للمرحلة الثانوية
الأفعال وأنواعها	نصب الفعل المضارع للمرحلة الثانوية
الإعراب والبناء	إعراب المثنى والجموع بأنواعها
الإعراب والبناء	الإعراب والبناء
الإعراب والبناء	الجمل التي لها (وليس لها) محل من الإعراب
الإعراب والبناء	المعرب والمبني من الأسماء
الإعراب والبناء	المعرب والمبني من الأفعال
الإعراب والبناء	علامات إعراب المضاف إليه
التوابع	البدل
التوابع	التوابع
التوابع	التوكيد
التوابع	العطف
التوابع	النعت المفرد، وغير المفرد
الإملاء والمعجم	الكشف في المعجم
الإملاء والمعجم	همزة القطع، وألف الوصل

Table 8: Taxonomy of grammar lessons (continued)

الملحق بالمثنى	الصرف والميزان
الملحق بجمع المؤنث السالم	الصرف والميزان
الملحق بجمع المذكر السالم	الصرف والميزان
الممنوع من الصرف	الصرف والميزان
الميزان الصرفي	الصرف والميزان
تثنية المقصور وجمعه	الصرف والميزان
تثنية الممدود وجمعه	الصرف والميزان
تثنية المنقوص وجمعه	الصرف والميزان
أحوال المبتدأ والخبر	المرفوعات والمنصوبات
أنواع الخبر	المرفوعات والمنصوبات
الأسماء الخمسة	المرفوعات والمنصوبات
الاستثناء	المرفوعات والمنصوبات
التمييز	المرفوعات والمنصوبات
الحال المفردة، والحال غير المفردة	المرفوعات والمنصوبات
المرفوعات	المرفوعات والمنصوبات
المفعول المطلق	المرفوعات والمنصوبات
المفعول لأجله	المرفوعات والمنصوبات
المفعول معه وحروف الجر الزائدة	المرفوعات والمنصوبات
المنادى	المرفوعات والمنصوبات
المنصوبات	المرفوعات والمنصوبات
النائب عن المفعول المطلق	المرفوعات والمنصوبات
علامات إعراب الفاعل والمبتدأ والخبر	المرفوعات والمنصوبات
علامات إعراب المفعول به	المرفوعات والمنصوبات
إعمال اسم الفاعل	المشتقات والمصادر
إعمال اسم المفعول	المشتقات والمصادر
إعمال صيغ المبالغة	المشتقات والمصادر
اسم الآلة	المشتقات والمصادر
اسم التفضيل	المشتقات والمصادر
اسم الفاعل	المشتقات والمصادر
اسم المرة واسم الهيئة	المشتقات والمصادر
اسم المفعول	المشتقات والمصادر
اسما الزمان والمكان	المشتقات والمصادر
المشتقّات	المشتقات والمصادر
المصادر	المشتقات والمصادر
المصدر الميمي والمصدر الصناعي	المشتقات والمصادر
صيغ المبالغة	المشتقات والمصادر
إن وأخواتها	النواسخ
النواسخ	النواسخ
كاد وأخواتها	النواسخ
كان وأخواتها	النواسخ

C Benchmarking Prompts

C.1 GU Benchmarking Prompt

GU Prompt

هذا سؤال في النحو العربي مع أربع إجابات. اختر الإجابة الصحيحة. اكتب حرف الإجابة فقط دون أي نص إضافي.
{{question}}

الخيارات:
A. {{A}}
B. {{B}}
C. {{C}}
D. {{D}}

الجواب:

Translation: This is a question in Arabic grammar with four candidate answers. Choose the correct answer. Write only the letter of the answer without any additional text.

C.2 GED Benchmarking Prompt

GED Prompt

في النص التالي أخطاء نحوية. مهمتك هي تحديد الكلمات التي فيها خطأ وإعادة قائمة بهذه الكلمات فقط ولا شيء آخر.
تأكد أن تعيد قائمة الكلمات التي فيها أخطاء بالصيغة التالية:
##الأخطاء: خطأ ١، خطأ ٢، خطأ ٣، ... ##
ولا تزد بعد ذلك أي شيء إضافي، لا تصحيح الكلمات ولا تعليل الخطأ.

مثال:

النص: ذهب الولدين إلا المدرسة

##الأخطاء: الولدين، إلا ##

النص:

Translation: There are grammatical errors in the following text. Your task is to identify the erroneous words and list only those words and nothing else. Make sure to list the erroneous words in the following format:
##Errors: error 1, error 2, error 3, ... ##
Do not add anything further, neither correct the words nor explain the error.

C.3 GEC Benchmarking Prompt

GEC Prompt

في النص التالي كلمة محددة بين قوسين ( ) فيها خطأ نحوي.
مهمتك هي اقتراح كلمة واحدة بديلة صحيحة نحويا مع إعطاء شرح للتصحيح. تأكد أن تعيد التصحيح بالصيغة التالية:
الخطأ: الكلمة الخاطئة، التصحيح: الكلمة الصحيحة، الشرح: شرح التصحيح

مثال:

النص: ذهب (الولدين) إلى المدرسة
الخطأ: الولدين، التصحيح: الولدان، الشرح: مثنى مرفوع وعلامة رفعه الألف

النص:

Translation: In the following text, a word marked in brackets ( ) contains a grammatical error. Your task is to suggest one grammatically correct alternative word, along with an explanation of the correction. Make sure to repeat the correction in the following format: ##Error: Erroneous word, Correction: Correct word, Explanation: Explanation of correction ##

C.4 GEX Benchmarking Prompt

The following prompt template was used to benchmark LLMs for the grammatical error explanation Task on *Nahw*-MCQ.

Nahw-MCQ GEX Prompt

```
You are an expert in Arabic grammar. You will be given a multiple-choice question with four options (A, B, C, D) and the correct answer.

### Instructions:
1. Read the question, the four answer choices, and the correct answer.
2. Provide a clear and concise explanation of why the correct answer is correct.
3. Do not restate the question or list the options.
4. Do not include any JSON or extra formatting. Only output the explanation text.

### Example:
Input:

الدرس: الأسماء الخمسة
الصف: ٧
السؤال: أغلِق ...... عند الغضب.
أكمل باسمٍ من الأسماء الخمسة
A) فاكَ
B) فوكَ
C) فيكَ
D) فمُكَ (بضمِ الميم)
الجواب الصحيح: A
Output:
كلمة (فاكَ) تعرب مفعولًا به منصوبًا، وعلامة نصبه الألف لأنه من الأسماء الخمسة.

—
### Input:
{question_text}

Output:
```

The following prompt template was used to benchmark LLMs for the grammatical error explanation task on *Nahw*-Passage.

Nahw-Passage GEX Prompt

```
In the following text, there is a specific word between curly brackets { } that contains a grammatical error. This is followed by the corrected text. Your task is to provide a short explanation of the grammatical rule behind the proposed correction.

##Example:
Text: ذهب {الولدين} إلى المدرسة
Correction: ذهب {الولدان} إلى المدرسة
Explanation: مثنى مرفوع وعلامة رفعه الألف


(Translation)
Text: The (Two) boys [genitive] went to school.
Correction: The (Two) boys [nominative] went to school.
Explanation: Dual Noun in nominative case; its marker is the Aif (ا).


Text: {text}
Correction: {correction}
Explanation:
```

D GEX Scoring Rubric

The model explanation was scored from 0 to 10 based on the scoring criteria in Table 9:

Score	Meaning
0	The explanation is completely absent or incorrect.
1–2	The correction is correct, but the explanation is absent or incorrect.
3–4	The explanation is present but very incomplete (it merely states that the word is incorrect without explaining why)
5–6	Average explanation: Identifies the error but lacks depth or generalization of the rule.
7–8	Good explanation: Identifies the error clearly and states a correct rule, with some deficiencies in detail or formulation.
9	Excellent explanation: Correct, precise, and explained with clear grammatical rules and appropriate context.
10	Complete explanation: Completely correct + clear rules + excellent pedagogical formulation + additional examples where needed.

Table 9: Scoring rubric for GEX

E Synthetic Data Generation

E.1 Prompt

The following prompt template was used to call GPT-4o (api_version: 2025-01-01-preview) to generate synthetic MCQs.

Synthetic Data Gen. Prompt

(Translation)
You are an expert in designing grammar tests for students across different grade levels. I want you to generate 10 multiple-choice questions in the form of a JSON Array, where all the questions come from a specific lesson and have a specified difficulty level.
### Instructions:

1. Generate exactly 10 questions.

2. All questions must be from the lesson: {lesson_name}

3. All questions must have the difficulty level: {difficulty}

4. The grade level for all questions must be: {grade_level}

5. For each question, create a JSON object containing exactly the following fields:
* "Grade": The grade level (must equal {grade_level}).
* "Lesson": The grammar lesson name (must equal {lesson_name}).
* "Difficulty": {difficulty}.
* "Question": The text of the question.
* "Choices": A list of choices (A, B, C, D) with the text of each option.
* "Answer": The letter of the correct option (A, B, C, D).
* "Explanation": An explanation clarifying why the answer is correct.
* "Question_Type": The question type (theoretical or applied).

6. Return the result as a JSON Array consisting of 10 elements, with no extra text or explanation outside the JSON structure.
### Example Format:

{examples}

E.2 Generation Error Analysis

Error Type	Explanation	%
Wrong Answer	Answer has errors or incorrect	32
Answers Count	More than one valid answer	29
Wrong Question	Question has no meaning	24
Wrong Difficulty	Easy questions labeled as hard	13
Repeated Answers	Answers are not unique	3

Table 10: Error analysis of the synthetic generated questions by GPT-4o

F Fine-tuning Setup and Hyperparameters

Table 11 details the specifications of the machine used to fine-tune the Gemma-3-4b-it models. Training was performed using Llamafactory (version 0.9.4.dev0) (Zheng et al., 2024). The hyperparameters employed during fine-tuning are summarized in Table 12. These values were chosen based on empirical experimentation, as they produced the best-performing models across different configurations of learning rate, weight decay, and batch size.

CPU	Intel(R) Xeon(R) Platinum 8568Y+
GPU	Nvidia H200 - 141GB
Memory	1TB
OS	Ubuntu 22.04.2 LTS

Table 11: Specification of the machine used to train the Gemma-3-4b-it models

Parameter	Value
Batch Size	32
Learning Rate	5e-6
Weight Decay	0.1
ADAM Beta1	0.9
ADAM Beta2	0.95
lr Scheduler Type	cosine
Warmup Ratio	0.06
Cutoff Len	2048

Table 12: Training parameters for Gemma-3-4b-it fine-tuning