

# To Ask LLMs about English Grammaticality, Prompt Them in a Different Language

**Shabnam Behzad**  
Georgetown University  
shabnam@cs.georgetown.edu

**Amir Zeldes**  
Georgetown University

**Nathan Schneider**  
Georgetown University

## Abstract

In addition to asking questions about facts in the world, some internet users—in particular, second language learners—ask questions about language itself. Depending on their proficiency level and audience, they may pose these questions in an L1 (first language) or an L2 (second language). We investigate how multilingual LLMs perform at *crosslingual* metalinguistic question answering. Focusing on binary questions about sentence grammaticality constructed from error-annotated learner corpora, we prompt three LLMs (Aya, Llama, and GPT) in multiple languages, including English, German, Korean, Russian, and Ukrainian. Our study reveals that the language of the prompt can significantly affect model performance, and despite English being the dominant training language for all three models, prompting in a different language with questions about English often yields better results.

## 1 Introduction

Mastering a second language is a skill acquired over time. Writing or speaking with proper grammar in a second language can be challenging, primarily due to the complexities of form, meaning, and the relationship between the two (DeKeyser, 2005). Learners often become confused when constructing sentences and seek help from teachers, peers, or online resources for clarity. Many of these queries relate to grammar; in fact, the *grammar* tag is the most popular one on online question-and-answer forums like English Language Learners StackExchange.<sup>1</sup>

Behzad et al. (2023) collected ELQA, a dataset of English questions and answers from such forums and their study showed that around 38% of the questions are *Fluency* questions. A common format within the Fluency category involves com-

paring multiple sentences to determine which one is correct:

**Example 1** *Which one is grammatically correct?*  
*He is believed to be awarded the prize at the age of 17 when he was in London.*

*or*

*He is believed to have been awarded the prize at the age of 17 when he was in London.*<sup>2</sup>

Behzad et al. (2023) investigate the capabilities of large language models (LLMs) in responding to learner questions including fluency and grammar questions and demonstrate that, although the performance has not yet reached human levels, the answers are accurate in many instances. This study was limited to asking questions in English and did not take into account that some learners may prefer to interact with LLMs in their native language, especially when they are at a novice level of proficiency. Moreover, Behzad et al.’s (2023) study only included questions about the English language. This motivated us to study the capabilities of LLMs in a multilingual setting as a first step towards building better interactive multilingual NLP technologies that can help learners practice a new language and answer their questions.

Since learner question answering datasets similar to ELQA are not available in other languages, and since we also wanted to ensure a controlled multilingual comparison, we synthetically created them by filling a template with naturally occurring sentences written by L2 learners of different languages. These sentences are selected and processed from grammatical error correction (GEC) datasets, ensuring they follow the same error patterns as those made by learners in the real world.

Using this template format, we experiment with multiple LLMs, comparing their performance across different languages from several perspec-

<sup>1</sup><https://ell.stackexchange.com/tags>

<sup>2</sup><https://ell.stackexchange.com/questions/151043/is-believed-to-grammar-sentence-which-one-is-correct>

tives. In summary, we answer the following research questions in this paper:

- **Performance beyond English:** Do LLMs perform better when questions are asked in and about English, which is usually the predominant language in the training data? Or is the performance comparable across other languages?
- **Crosslingual performance:** What is the impact of the language in which the question is framed vs. the language the question is about? Does code-switching (question/prompt in one language, and examples in a different language) affect performance?
- **Language selection:** Given that we can frame the prompt in several different languages, which ones can help improve grammar QA in a lower-resourced language?
- **Models' strengths and weaknesses:** Do models perform better on specific error types? Does the learner's proficiency level affect the model performance? How do models explain incorrect predictions?

## 2 Related Work

As large language models (LLMs) impact communities worldwide, researchers are increasingly studying their behavior across different languages and tasks (Winata et al., 2022; Michaelov et al., 2023; Qin et al., 2024; Shen et al., 2024; Dandaru et al., 2024). Ahuja et al. (2023) introduced MEGA for benchmarking LLMs, covering 16 NLP datasets across 70 typologically diverse languages. Tasks include classification, question answering, sequence labeling, generation (summarization), and responsible AI. The authors show that there is a consistent performance gap between high-resource, Latin script, and low-resourced languages. Lai et al. (2023) studied ChatGPT's behavior in a multilingual setting for several tasks (PoS tagging, named entity recognition, relation extraction, natural language inference, question answering, common sense reasoning, and summarization) and they argue that English task descriptions for non-English inputs can be better processed by ChatGPT, although other studies argue otherwise (Hasan et al., 2024; Deng et al., 2024; Liu et al., 2024).

Our focus in this paper is grammatical question answering for L2 language learners. A similar task to the task studied in this paper is the Grammatical Error Correction (GEC) task in which the goal is to

automatically identify and correct errors in a given text. The field has made remarkable progress over the past decade, largely inspired by multiple shared tasks introducing new benchmarks and agreed-upon standards for evaluating GEC systems (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013, 2014; Bryant et al., 2019; Volodina et al., 2023). So far, core approaches to GEC include rule-based, classification and classic machine learning approaches, statistical and neural machine translation, and edit-based approaches (Bryant et al., 2023; Wang et al., 2021).

More recently, the growing prominence of LLMs has led researchers to increasingly use them as zero-shot or few-shot generators for the error correction task in English (Loem et al., 2023) or other languages such as Chinese (Fan et al., 2023; Li et al., 2024), Arabic (Kwon et al., 2023) and Korean (Maeng et al., 2023).

Most of these studies focus on GPT family models and many have come to the conclusion that the results are not yet satisfactory because even though output sentences are fluent and grammatical, LLMs tend to over-correct sentences as they go beyond minimal edits (Sottana et al., 2023; Wu et al., 2023; Fang et al., 2023; Davis et al., 2024). Katinskaia and Yangarber (2024) reveals that for multiple languages, including Czech, German, Russian, Spanish, and Ukrainian, GPT makes significant alterations to the source sentences, and sometimes even changes their semantics which creates major challenges for evaluation with reference-based metrics. Automatic evaluation in text generation has been a challenge for some time (Celikyilmaz et al., 2020; Gehrmann et al., 2023), and GEC is no exception (Rozovskaya and Roth, 2021; Gong et al., 2022; Östling et al., 2024).

## 3 Proposed Study

To evaluate LLMs' multilingual capabilities in L2 learner question answering, we create a binary question answering template that is easily adaptable to different languages. The template is similar to patterns observed by Behzad et al. (2023) and shown in Example 1. The body of the question in English is *Which sentence is more grammatical and native-like?* This can be translated to any other language, such as German: *Welcher Satz ist grammatikalischer und muttersprachlicher?* Therefore, we can extend the template to encompass other languages. The question body comprises two

Template	Data Instance	Correct Answer
in eng about eng	Which sentence is more grammatical and native-like? 1) It helps me healthy . 2) <b>It helps me keep healthy .</b>	2
in eng about deu	Which sentence is more grammatical and native-like? 1) Fühlen sich etwas orientierungslos . 2) <b>Sie fühlen sich etwas orientierungslos .</b> (translation: <i>You feel a bit disorientated.</i> )	2
in kor about deu	어떤 문장이 더 문법적이고 자연스러운가요? 1) <b>Das nützt uns für unseren Beruf .</b> (translation: <i>This is useful for our profession.</i> ) 2) Das nützt für unseren Beruf .	1

**Table 1:** Examples instances from our evaluation benchmark for different languages. The correct sentence is shown in green. English translations are shown for clarity (they were not provided to the system).

sentences: one grammatical and one containing an error. The order is randomized to establish 50% as the accuracy expected due to chance. This template format also allows us to avoid the challenges associated with reference-based text generation evaluation.

While simple, this step is crucial for developing question-answering systems for learners. Before generating any type of explanation for the learner, we must first confirm the model’s capability to differentiate between erroneous and grammatical sentences.

To ensure experimentation with learner-like data, we incorporate sentence pairs from GEC benchmarks into our template. This setup maintains high-quality data, and simplifies the process of switching between different languages, both for the question itself and for the example sentences. Examples of our benchmark instances are available in Table 1.

### 3.1 Data

We focus on English, German, Russian, Ukrainian, and Korean in this paper, for which human-annotated GEC datasets already exist. We wanted to include multiple scripts in our experiments (Cyrillic, Latin, and Hangul) and we specifically included Ukrainian since it is a mid-resource language compared to the rest of the languages studied in this paper which are considered high-resource (Ustun et al., 2024; Joshi et al., 2020).

**English** We used the W&I dataset (Yannakoudakis et al., 2018) which is collected from an online platform used by non-native English students and manually annotated for correction and CEFR (Common European Framework of Reference for Languages) levels.

**German** We used the data from Falko (Reznicek et al., 2012) which presented annotation guidelines

for minimal target hypotheses, and then used these for annotations and assessment of German learner essays. The data was later processed and made available in the M2 format by Boyd (2018) for the GEC task.

**Russian** We obtained the RULEC (Alsufovea et al., 2012) dataset. The corpus comprises essays and papers written by students learning Russian as a foreign language and heritage speakers in a university setting in the United States. Later, Rozovskaya and Roth (2019) added manual corrections to a subset of this corpus and made it available for the GEC task.

**Ukrainian** We used UA-GEC (Syvokon et al., 2023). The dataset encompasses a diverse range of writing genres, including text chats, essays, and formal writing. Human annotators corrected and annotated the corpus for errors related to fluency, grammar, punctuation, and spelling.

**Korean** We used data collected by Yoon et al. (2023). The dataset includes data from native Korean speakers, Korean as a foreign language learners which was corrected by Korean tutors, and KorLang8 which was corrected by native Koreans on social platforms. We used the subset written by L2 learners (Kor-Learner) and corrected by Korean tutors.

When a learner asks a question, it usually focuses on a specific grammatical concept. Hence, we wanted the erroneous sentence in our template to only contain one error. This setup also makes the task more challenging since a sentence with multiple grammatical errors would be easier to identify. All these datasets included sentence-level data in the M2 format,<sup>3</sup> so we were able to update sen-

<sup>3</sup>This format includes the original sentence, as well as correction annotations which consists of the start and end token offset of the edit, the error type, and the tokenized

tences with multiple errors to incorporate corrections for all but one error. We removed *spelling* and *punctuation* errors and further made sure the edit distance between the original sentence and the corrected version is at least 4 characters.

Each test set comprises a random subset of 250 samples from our datasets, with labels 1 and 2 (indicating whether the first or second sentence is correct) balanced within each benchmark.

We experiment with few-shot learning (Brown et al., 2020), and include 4 examples of the task (from the train set) at inference time.

### 3.2 Language Models

We experimented with two open-source models and one closed model. For the open-source models, we used two different sizes: Aya with 13 billion parameters and Llama with 8 billion parameters.

**GPT** GPT-4 (Achiam et al., 2023) has demonstrated remarkable performance across numerous tasks and claims to be effective in various languages. We included this model in our study to compare its performance with the open-source models. However, it is important to note that the performance of GPT-4 is likely overestimated, as the model has probably already been exposed to the data we used. For our experiments, we used gpt-4-0125-preview.

**Aya** aya-101<sup>4</sup> (Ustun et al., 2024) is one of the few open-sourced, massively multilingual generative language models, covering up to 101 languages. It outperforms mT0 and BLOOMZ (Muennighoff et al., 2023) on most tasks while supporting twice as many languages.

**Llama** We use Meta-Llama-3-8B-Instruct<sup>5</sup> in our experiments. The Llama 3 (AI@Meta, 2024) instruction-tuned models are specifically optimized for dialogue applications and exceed the performance of many open-source chat models on standard industry benchmarks. While Llama is primarily intended for English use, its training data includes information from over 30 languages.<sup>6</sup>

Although the grammatical correction capabilities of GPT have been studied in previous works, we believe we are among the first to examine the capabilities of Llama and Aya in this domain and compare them to GPT.

corrected string.

<sup>4</sup><https://huggingface.co/CohereForAI/aya-101>

<sup>5</sup><https://huggingface.co/meta-llama/>

Meta-Llama-3-8B-Instruct

<sup>6</sup><https://ai.meta.com/blog/meta-llama-3/>

Setup	Llama	Aya	GPT
in eng about eng	72.0	<b>50.8</b>	67.6
in deu about eng	<b>80.0</b>	44.4	75.2
in rus about eng	76.0	44.8	77.6
in ukr about eng	76.8	38.8	<b>79.6</b>
in kor about eng	74.8	47.2	72.0

**Table 2:** Performance (accuracy) on 250 English sentence pairs, when questions are asked in different languages

Setup	Llama	GPT
in ukr about deu	73.2	<b>82.4</b>
in deu about deu	<b>73.6</b>	78.8
in ukr about rus	<b>74.0</b>	78.0
in deu about rus	70.0	<b>78.4</b>
in rus about rus	73.2	77.6
in ukr about ukr	56.0	<b>60.0</b>
in deu about ukr	<b>67.6</b>	50.0
in ukr about kor	<b>69.6</b>	<b>64.8</b>
in deu about kor	67.6	50.0
in kor about kor	65.2	42.8

**Table 3:** Comparing model accuracy when the question is asked in the same language as sentence pairs, vs. when it is asked in another language and there is code-switching

## 4 Evaluation and Results

In this section, we report and discuss our findings. Examining the results from experimenting with all combinations of languages, we observe the highest accuracy for grammar questions about English is 80% using Llama; for questions about German it is 84% using GPT; Korean, 69.6% using Llama; Russian, 78.4% using GPT; and Ukrainian, 62.4% using GPT (results for all different combinations are available in Appendix A, Table 11). These results reveal a performance gap between different languages on this task, indicating that there is still significant room for improvement in grammatical error detection in multilingual settings.

Next, we answer the following research questions based on our empirical results:

**Are LLMs consistent in performance when the same question is asked in different languages?**  
No.

To evaluate this, we keep English sentence pairs constant, but translate the first part of the prompt (“Which sentence is more grammatical and native-like?”) into different languages. Since the knowledge required to answer the question remains the same (English grammar) we expect language models to perform consistently among different languages. However, results in Table 2 show otherwise:

First, we note that Aya’s performance is below



random chance (all test benchmarks are balanced) in all languages except for English, where it is roughly at chance.

Llama performs fairly well in all languages, getting the highest accuracy (80%) when the question is asked in German, which is 8 percentage points higher than when it is asked in English.

GPT also performs better when the question is asked in a different language. There is a 12% increase in accuracy when question is asked in Ukrainian instead of English.

This is particularly interesting since, for both models, English is the dominant language in the training data. We hypothesize that since the model is heavily fine-tuned for a variety of English tasks that are not metalinguistic, the English prompts are less likely to be interpreted metalinguistically. So when the model encounters different languages, it anticipates metalinguistic usage and also focuses more on sentence structure. Another possible explanation could be that these models undergo extensive instruction fine-tuning specifically for English tasks. This intensive fine-tuning may potentially impair some of the model’s broader language modeling capabilities, particularly within English contexts. In contrast, the performance observed in other languages may reflect the advantage of less instruction fine-tuning, especially for less common tasks.

### Do LLMs perform better when questions are asked in specific languages? Yes.

In Table 2 we observed that GPT and Llama perform better in English grammar error detection when the question is asked in Ukrainian and German, respectively. However, is this pattern consistent across other languages? To investigate this, we expanded our experiments to include setups where questions are asked in the same language as the sentence pairs, as well as in German and Ukrainian. We present these results in Table 3.

Except for questions about Russian grammar, GPT seems to perform better when the question is asked in Ukrainian rather than in the same language as the sentence pair. We hypothesize that the reason *in deu about rus* gives better performance than *in ukr about rus* is because Ukrainian and Russian both use the Cyrillic script, whereas a different script such as Latin may help the model’s attention. It is interesting to note that not all different scripts work similarly; for instance, GPT achieves 70% accuracy in the *in kor about rus* setup. We

Setup	Llama	Aya	GPT
in ukr about ukr	56.0	46.0	60.0
in eng about ukr	55.6	48.4	60.0
in deu about ukr	58.4	50.8	54.8
in rus about ukr	60.0	49.2	62.4
in kor about ukr	57.2	45.6	49.2

**Table 4:** Studying Ukrainian as a mid-resource language: does asking the question in a high-resource language improve performance? Red and blue coloring indicate the lowest and highest scores for each language model, respectively.

performed McNemar’s test to make sure the differences were significant. The  $p$ -value for (*in deu about rus*, *in ukr about rus*) is  $9.86e-19$ , (*in kor about rus*, *in ukr about rus*) is  $4.27e-15$  and for (*in deu about rus*, *in kor about rus*) is  $3.11e-13$  which shows that differences are significant and that possibly the prompt being in a different script but belonging to the same family (Indo-European) might be beneficial to the model’s performance.

With Llama, it appears that code-switching generally improves performance, particularly with Ukrainian grammar, which is considered a mid-resource language. There is an 11.6% performance boost when questions are asked in German rather than in Ukrainian.

### Does prompting in a high-resource language aid grammatical error detection in a lower-resource language? Yes (for Ukrainian).

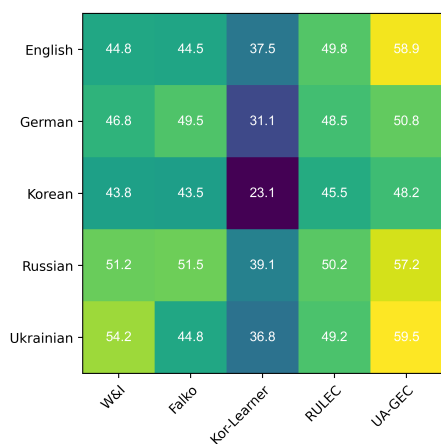
In Table 4 we present results on asking questions about Ukrainian grammar in different languages. Surprisingly, asking questions in English does not yield the best performance.

GPT achieves its highest performance when the question is asked in Russian. Similarly, Llama also performs best when the question is asked in Russian. These findings suggest that in studies involving grammar knowledge in a mid-resource language and when models are less finetuned on a language, prompting in a high-resource language with the same script can enhance performance.

We focused solely on Ukrainian in this study and, therefore, do not intend to make broad claims at this stage. We believe this hypothesis should be validated across different languages before drawing any major conclusions.

### Are the same sentence pairs equally challenging for different models? For some datasets, yes.

For this analysis, we compare the predictions of Llama and GPT and compute agreement percentage. Results can be found in Figure 1.



**Figure 1:** Prediction agreement percentage between Llama and GPT in different experimental setups. The lowest agreement (23.1%) is when questions are in and about Korean. The highest agreement (59.5%) is when questions are in and about Ukrainian.

First looking at Kor-Learner, we observe that, regardless of the language of the prompt, models exhibit the most disagreements on Korean grammar, highlighting distinct strengths and weaknesses between Llama and GPT in this language.

Next, we observe that the two models have the least number of disagreements on Ukrainian grammar. Looking back at Table 4, both models’ accuracy is below 63% in all different setups, which means both models often predict the wrong label and there is much room for improvement for this task in lower resourced languages.

**What if we ask the model to explain a wrong prediction?** The performance will improve.

Chain-of-Thought Prompting (Wei et al., 2022) has improved performance on tasks like math problems and common sense reasoning. Prior work shows that debating or providing feedback to LLMs can enhance their performance (Madaan et al., 2023; Wang et al., 2023; Pan et al., 2024). Here, we investigate with a small-scale experiment how models explain an incorrect prediction, and whether they would change their answer when asked for an explanation.

We experiment with both Llama and GPT, focusing on the *in eng about eng* setup due to budgetary constraints. If the model selects the incorrect sentence, we follow up with the question: *Can you explain why?*

One of the authors manually reviewed 50 examples from each model. Both models corrected their responses in more than 60% of cases. For the remaining cases, two main issues were ob-

served: 1) Both sentences were correct but conveyed slightly different meanings (30% of cases. Example: first row in Table 5). Corrections in the GEC dataset might have been made because the tutor had access to more context in the essay, or due to a personal preference for specific usage. 2) In a small number of cases (less than 10%), the models provided incorrect or irrelevant explanations, with this occurring more frequently in Llama responses (such as the second row in Table 5). These results suggest that engaging in conversation and asking follow-up questions when using LLMs can be helpful. However, it may also confuse learners if the model frequently changes its response.

## 5 Analysis and Discussion

In this section, we take a look at cases where LLMs struggle. Mainly, we study this from two perspectives: i) the proficiency level of the learner who wrote the sentence and ii) the type of error that exists in the sentence.

### 5.1 Learner’s Proficiency Level

Second language learners at different proficiency levels make different types of errors (Zeng et al., 2024). Here, we investigate whether models perform better on sentences written by learners with specific proficiency levels. The proficiency level of the learner is only available for the English and German datasets. For English, we have corresponding metadata for CEFR levels, while for German, we utilized C-test scores of the learners, which we mapped to CEFR levels (Jaworska et al., 2015).

Results on W&I with no code-switching are available in Table 6. Aya hardly achieves above-chance performance on any of the proficiency levels. For Llama, performance ranges from 66% on beginner-level sentences to 79% on advanced-level sentences. GPT exhibits a similar trend to Llama, with its lowest performance (67.0%) on beginner-level sentences and its highest performance (71.2%) on advanced-level sentences. One would expect the model to perform better on beginner-level errors. However, looking at a subset of examples manually, we found some beginner-level sentences difficult to follow and challenging to correct without additional context.

Results on Falko are available in Table 7. Again, Aya’s performance is not acceptable as it is below chance level. The performance patterns of Llama and GPT align with the results on W&I: Llama

Model	Sentence Pair and Explanation
GPT	<p>1) For the above mentioned reasons , I am going to conclude that travelling by car is much more convenient .</p> <p>2) Due to the above mentioned reasons , I am going to conclude that travelling by car is much more convenient .</p> <p>Certainly! Both sentences are grammatically correct and convey the message clearly, but there are subtle differences in their structure and the way they might be perceived by native speakers.</p> <p>1. "For the above mentioned reasons, I am going to conclude that travelling by car is much more convenient." - This sentence is correct and understandable. The phrase "For the above mentioned reasons" serves as a prepositional phrase that introduces the basis for the conclusion. However, it might feel slightly more formal or less fluid in casual conversation.</p> <p>2. "Due to the above mentioned reasons, I am going to conclude that travelling by car is much more convenient." - This sentence also correctly conveys the message. The phrase "Due to the above mentioned reasons" is commonly used in both written and spoken English to indicate causality or the reason for something. It flows well and might be perceived as slightly more natural or smoother in everyday language.</p> <p>The preference for the second sentence stems from its use of "Due to," which is a common way to introduce reasons or causes in English, making it sound more native-like in this context. It succinctly links the reasons to the conclusion, which is why it might be preferred for its clarity and fluency in communication.</p>
Llama	<p>1) Finally concert started and all the lights were shining like stars .</p> <p>2) Finally the concert started and all the lights were shining like stars .</p> <p>In English, when we use the phrase "finally" to indicate that something has happened after a long delay or anticipation, it typically takes the subject-verb-object (SVO) word order. This means that the verb "started" should come after the word "finally".</p> <p>In sentence 1, "Finally concert started" is not grammatically correct because "concert" is the subject, and "started" is the verb. The word order is not following the typical SVO pattern.</p> <p>In sentence 2, "Finally the concert started" is grammatically correct because the word "the concert" is the subject, and "started" is the verb, following the typical SVO pattern.</p> <p>Additionally, sentence 2 sounds more native-like because it follows the typical word order and phrasing that native English speakers would use. Sentence 1, on the other hand, has a slightly unusual word order that might make it sound less natural to native English speakers.</p> <p>So, the correct answer is sentence 2: "Finally the concert started and all the lights were shining like stars."</p>

**Table 5:** Examples of explanations provided by models when the incorrect sentence was selected by them. The correct sentences in the first and second rows are the first and second sentences respectively.

Model	A	B	C
Llama	66.0	77.0	78.8
Aya	55.0	45.0	53.8
GPT	67.0	71.0	71.2

**Table 6:** Model performance (accuracy) on W&I (English) dataset with respect to the proficiency level of the learner. A (beginner), B (intermediate), C (advanced).

Model	B	C
Llama	70.9	78.0
Aya	45.6	44.7
GPT	76.7	74.7

**Table 7:** Model performance (accuracy) on Falko (German) dataset with respect to the proficiency level of the learner. B (intermediate), C (advanced).

performs better on intermediate-level sentences but struggles more with advanced-level sentences, while GPT’s performance is fairly similar for both intermediate and advanced errors.

## 5.2 Grammatical Error Types

For datasets where we have fine-grained annotation on error types, we examine model performance

Error Types	Aya	Llama	GPT
DET	44/77	53/77	42/77
NOUN	23/42	31/42	30/42
PREP	14/35	22/35	26/35
VERB	24/46	40/46	33/46
VERB:TENSE	23/38	34/38	29/38

**Table 8:** The fraction of the number of correct predictions over the total number of samples in an error category on the W&I benchmark (in eng about eng) using universal POS tags (Petrov et al., 2012). Only error categories with at least 20 samples in our benchmark were included.

across different error categories to determine if certain types are more challenging than others.

**English** Results on the templates from W&I are reported in Table 8. We observe that errors categorized under verb tense are easier for all three models, with Llama achieving 89.5% accuracy in this category, surpassing GPT (76.3%). The most challenging category for GPT is determiner-related errors, where it performs near chance level. For both Aya and Llama, preposition-related errors pose significant challenges.

Error Types	Aya	Llama	GPT
ADP	13/32	27/32	27/32
ADV	9/11	11/21	12/21
AUX	21/45	29/45	38/45
DET	27/66	53/66	44/66
PRON	15/31	24/31	24/31
VERB	14/28	22/28	26/28
WO	13/26	20/26	21/26

**Table 9:** The fraction of the number of correct predictions over the total number of samples in an error category on the Falko benchmark (in deu about deu) using universal POS tags (Petrov et al., 2012); WO stands for word-order. Only error categories with at least 20 samples in our benchmark were included.

Error Types	Aya	Llama	GPT
Case	16/35	15/35	22/35
CONJ	15/25	14/25	15/25
PREP	11/27	17/21	11/27
Ungrammatical Structure	43/95	63/95	66/95

**Table 10:** The fraction of the number of correct predictions over the total number of samples in an error category on the UA-GEC benchmark (in ukr about ukr). Only error categories with at least 20 samples in our benchmark were included.

**German** Results are available in Table 9. Models tend to make the most mistakes on German adverb related errors. Llama’s performance on this category is near chance level.

Errors related to the verb seem to be among the easier categories, with an accuracy of 93% for GPT and 79% for Llama. For Llama, detecting determiner errors seems to be the easiest; it achieves an accuracy of 80% on these.

**Ukrainian** We present results on UA-GEC in Table 10. One of the few setups where Aya achieves performance above chance level is with Ukrainian conjunction errors, matching GPT’s performance (60%). Llama performs below chance on case errors, while GPT struggles with preposition errors. Identifying ungrammatical structures (divergence from syntactic norms) appears to be relatively easier for both Llama and GPT, with accuracies of 66.3% and 69.5%, respectively.

Examining some of these sentences, the errors are relatively simple, yet the model still fails to select the correct sentence. This indicates that there is considerable room for improvement in error detection, let alone in generating explanations.

## 6 Conclusion

In this paper, we study the multilingual capabilities of several language models in a learner metalinguistic question answering task. Our results show

that i) there is still room for improvement in this task across different languages including English; ii) models are too sensitive to the prompt and the language of the prompt; and iii) asking questions in languages other than English can help with performance, even for questions about English grammar.

A particularly worrying result, which may be counter-intuitive to downstream users of LLMs, is the substantial degree of variance in how models respond to the same sentence pairs based purely on the language in which the question is posed. Here we see a clear trend favoring asking in a different language, with questions in a different script showing particularly notable gains, possibly because they sensitize the model to a meta-linguistic environment (i.e. asking questions about a sentence, rather than asking the model to respond to the contents of the sentence itself).

We hope the findings of this research may shed light on the challenges of multilingual, metalinguistic Q&A and enable future work on this task. That results vary widely across models—without a clear correlation with parameter count, language, or error type—suggests that more attention needs to be paid to metalinguistic Q&A in order to obtain satisfactory and consistent performance.

## 7 Limitations

Our findings are specific to the models we tested, and it’s possible that other models could perform better. However, validating this assumption exceeds our current scope and computational resources. We did not fine-tune the models, which could potentially enhance performance, nor did we explore various prompting strategies beyond few-shot and chain of thought. Additionally, each experiment was only done once. Running multiple experiments could give more reliable and consistent results.

Furthermore, the quality of our template and sentence pairs heavily relies on the GEC datasets used, and any potential errors in the original data would propagate to our experiments as well.

Lastly, our study focuses on five languages, leaving room for expanded inclusivity by testing additional languages and offering a more comprehensive understanding of the current state of the art and its limitations.

## Acknowledgements

We thank the anonymous reviewers for their insightful comments. This research was supported in



part by NSF award IIS-2144881, and a Georgetown GSAS-GradGov research grant.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Kabir Ahuja, Harshita Diddee, Rishav Hada, Milliecent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual evaluation of generative AI](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 model card](#).
- Anna Alsufieva, Olesya Kisselev, and Sandra Freels. 2012. [Results 2012: Using flagship data to develop a Russian learner corpus of academic writing](#). *Russian Language Journal*, 62:79–105.
- Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. [ELQA: A corpus of metalinguistic questions and answers about English](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2047, Toronto, Canada. Association for Computational Linguistics.
- Adriane Boyd. 2018. [Using Wikipedia edits in low resource grammatical error correction](#). In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. [The BEA-2019 shared task on grammatical error correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. [Grammatical Error Correction: A Survey of the State of the Art](#). *Computational Linguistics*, 49(3):643–701.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of text generation: A survey](#). *ArXiv*, abs/2006.14799.
- Robert Dale, Ilya Anisimoff, and George Narroway. 2012. [HOO 2012: A report on the preposition and determiner error correction shared task](#). In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.
- Robert Dale and Adam Kilgarriff. 2011. [Helping our own: The HOO 2011 pilot shared task](#). In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France. Association for Computational Linguistics.
- Christopher Davis, Andrew Caines, Øistein Andersen, Shiva Taslimipour, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. [Prompting open-source and commercial language models for grammatical error correction of english learner text](#). *arXiv preprint arXiv:2401.07702*.
- Robert M. DeKeyser. 2005. [What makes learning second-language grammar difficult? a review of issues](#). *Language Learning*, 55(S1):1–25.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. [Multilingual jailbreak challenges in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Ryandito Diandaru, Lucky Susanto, Zilu Tang, Ayu Purwarianti, and Derry Tanti Wijaya. 2024. [Could we have had better multilingual LLMs if English was not the central language?](#) In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024*, pages 43–52, Torino, Italia. ELRA and ICCL.
- Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. [Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning](#). In *Natural Language Processing and Chinese Computing*, pages 69–80, Cham. Springer Nature Switzerland.
- Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. [Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation](#). *ArXiv*, abs/2304.01746.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. [Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text](#). *J. Artif. Int. Res.*, 77.

- Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. [Revisiting grammatical error correction evaluation and beyond](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. [Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818, Torino, Italia. ELRA and ICCL.
- Sylvia Jaworska, Cédric Krummes, and Astrid Ensslin. 2015. [Formulaic sequences in native and non-native argumentative writing in german](#). *International Journal of Corpus Linguistics*, 20:500–525.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Anisia Katinskaia and Roman Yangarber. 2024. [GPT-3.5 for grammatical error correction](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.
- Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [Beyond English: Evaluating LLMs for Arabic grammatical error correction](#). In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.
- Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. [ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Yinghui Li, Shang Qin, Jingheng Ye, Shirong Ma, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2024. [Rethinking the roles of large language models in chinese grammatical error correction](#). *arXiv preprint arXiv:2402.11420*.
- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. [Is translation all you need? a study on solving multilingual tasks with large language models](#). *ArXiv*, abs/2403.10258.
- Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. [Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Junghwan Maeng, Jinghang Gu, and Sun-A Kim. 2023. [Effectiveness of ChatGPT in Korean grammatical error correction](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 464–472, Hong Kong, China. Association for Computational Linguistics.
- James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. [Structural priming demonstrates abstract grammatical representations in multilingual language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. [The CoNLL-2014 shared task on grammatical error correction](#). In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.
- Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. [The CoNLL-2013 shared task on grammatical error correction](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.
- Robert Östling, Katarina Gillholm, Murathan Kurfalı, Marie Mattson, and Mats Wirén. 2024. [Evaluation](#)

- of really good grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6582–6593, Torino, Italia. ELRA and ICCL.
- Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 11:484–506.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC 2012*, pages 2089–2096, Istanbul, Turkey.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.
- Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das falko-handbuch. korpusaufbau und annotationen. version 2.01.
- Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.
- Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online. Association for Computational Linguistics.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.
- Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.
- A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *ArXiv*, abs/2402.07827.
- Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.
- Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.
- Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Genta Winata, Shijie Wu, Mayank Kulkarni, Tamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.
- Hao Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael R. Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *ArXiv*, abs/2303.13648.
- Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.
- Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean grammatical error correction: Datasets and annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.
- Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. Evaluating prompting

strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6426–6430, Torino, Italia. ELRA and ICCL.



## A Experimental results

In the following table, we share our results for all different language combinations.

Setup	Llama	Aya	GPT
in eng about eng	72.0	50.8	67.6
in deu about eng	80.0	44.4	75.2
in rus about eng	76.0	44.8	77.6
in ukr about eng	76.8	38.8	79.6
in kor about eng	74.8	47.2	72.0
in eng about deu	68.8	43.6	77.2
in deu about deu	73.6	44.4	78.8
in rus about deu	69.2	50.8	84.0
in ukr about deu	73.2	45.2	82.4
in kor about deu	73.6	42.8	76.0
in eng about rus	66.8	49.6	74.8
in deu about rus	70.0	53.6	78.4
in rus about rus	73.2	45.6	77.6
in ukr about rus	74.0	42.8	78.0
in kor about rus	73.6	46.0	70.0
in eng about ukr	55.6	48.4	60.0
in deu about ukr	58.4	50.8	54.8
in rus about ukr	60.0	49.2	62.4
in ukr about ukr	56.0	46.0	60.0
in kor about ukr	57.2	45.6	49.2
in eng about kor	66.4	46.0	55.6
in deu about kor	67.6	46.0	50.0
in rus about kor	66.0	47.6	63.6
in ukr about kor	69.6	50.0	64.8
in kor about kor	65.2	46.4	42.8

**Table 11:** Performance (accuracy) on all different language combinations.