

GPT-3.5 for Grammatical Error Correction

Anisia Katinskaia,^{*◇} Roman Yangarber[◇]

^{*} Department of Computer Science, [◇] Department of Digital Humanities
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

This paper investigates the application of GPT-3.5 for Grammatical Error Correction (GEC) in multiple languages in several settings: zero-shot GEC, fine-tuning for GEC, and using GPT-3.5 to re-rank correction hypotheses generated by other GEC models. In the zero-shot setting, we conduct *automatic* evaluations of the corrections proposed by GPT-3.5 using several methods: estimating grammaticality with language models (LMs), the Scribendy test, and comparing the semantic embeddings of sentences. GPT-3.5 has a known tendency to over-correct erroneous sentences and propose alternative corrections. For several languages, such as Czech, German, Russian, Spanish, and Ukrainian, GPT-3.5 substantially alters the source sentences, including their semantics, which presents significant challenges for evaluation with reference-based metrics. For English, GPT-3.5 demonstrates high recall, generates fluent corrections, and generally preserves sentence semantics. However, *human* evaluation for both English and Russian reveals that, despite its strong error-detection capabilities, GPT-3.5 struggles with several error types, including punctuation mistakes, tense errors, syntactic dependencies between words, and lexical compatibility at the sentence level.

Keywords: GPT-3.5, LLM, GEC, grammatical error correction, evaluation

1. Introduction

Recent advances in large language models (LLM)—e.g., GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), Gemini (Team et al., 2023), Mixtral (Jiang et al., 2024), etc.—have resulted in strong performance on various NLP tasks in zero-shot or few-shot settings. Recent work (Coyne and Sakaguchi, 2023; Coyne et al., 2023; Fang et al., 2023; Loem et al., 2023) has assessed several GPT-3.5 models on the task of grammatical error correction (GEC), mostly for English, and compared with other state-of-the-art models, showing great potential of GPT-3.5 for GEC.

We experiment with multiple languages in several settings: zero-shot setting, including zero-shot chain-of-thought (CoT) (Kojima et al., 2022), fine-tuning GPT-3.5 models, and using GPT-3.5¹ to re-rank hypotheses generated by other GEC models. To our knowledge, GPT-3.5 has not been assessed for GEC in all three of these settings previously. Our experiments were performed at the sentence level using OpenAI’s GPT-3.5-turbo² models through the official OpenAI API.³

Since GPT-3.5 is known to suggest fluent corrections with high recall (Fang et al., 2023; Wu et al., 2023), reference-based metrics tend to underestimate the model’s true performance and provide only a partial estimation of it, as detailed in (Rozovskaya and Roth, 2021). However, we lack other metrics suitable for verifying that GPT-3.5 indeed corrects grammatical errors and does not rewrite

the input into a coherent yet irrelevant sentence. These increasing challenges in evaluating the performance of GEC models have motivated us to use manual and reference-free evaluation, as well as measuring semantic similarity between the input sentences and the corrected sentences.

We address the following research questions: (RQ1) Can GPT-3.5 be used consistently for GEC across different languages, in different settings, such as zero-shot and fine-tuning? (RQ2) Does GPT-3.5 alter the semantics of the source sentence while attempting to correct it? Based on our experiments, we make the following observations:

(A) In a zero-shot setting, GPT-3.5 achieves higher recall than other GEC models for the majority of the tested languages (except Czech and Ukrainian). However, as manual evaluation of sentences corrected by GPT-3.5 for English and Russian shows, corrected sentences can deviate significantly from the input. Fine-tuning performance is much lower for all tested languages, with numerous sentences remaining uncorrected.

(B) Using GPT-3.5 to re-rank correction hypotheses generated by other GEC models results in an increased recall, as the model ranks more fluent corrections higher. Re-ranking hypotheses with GPT-3.5 beats GPT-3.5’s performance in the zero-shot setting for GEC across all evaluated languages. This indicates that GPT-3.5 may be more effectively used as a re-ranker in scenarios where other smaller fine-tuned GEC models are available.

(C) A detailed manual evaluation of the corrections suggested by GPT-3.5 showed that the model is able to correct errors and to handle long sequences and reversed word order. However, it

¹<https://openai.com/blog/chatgpt>

²Version `gpt-3.5-turbo-0613`

³<https://platform.openai.com/docs/api-reference>

often struggles with correcting punctuation, article, preposition, and verb tense errors in English; it struggles with correcting errors in punctuation, government relations between words, and lexical compatibility in Russian.

(D) Reference-free evaluation of GPT-3.5 performance in a zero-shot setting—using pre-trained language models (LMs), Scribendy scores (Islam and Magnani, 2021), and evaluation of semantic similarity between sentence embeddings—reveal that GPT-3.5 corrections tend to deviate from the source sentences most in Czech, German, Spanish, and Ukrainian. The model modifies the input for these languages more liberally, which does not necessarily improve performance. For Russian, the results indicate that GPT-3.5 can correct more errors than any other GEC model, although its suggestions often semantically diverge from the source sentences. The performance for English appears more promising since the suggested corrections remain semantically similar to the input sentences. Additionally, over 50% of the sentences corrected by GPT-3.5 are evaluated by LMs as more fluent than the reference sentences. For Swedish, GPT-3.5 exhibits relatively high precision compared to other languages, with its correction hypotheses staying semantically close to the input, which may, though, result from under-corrected errors.

The paper is structured as follows. Section 2 provides a brief overview of related work on GEC. Section 3 details the datasets and describes experiments conducted in three settings: zero-shot, fine-tuning, and re-ranking of correction hypotheses. Section 4 discusses the results of these experiments. Section 5 focuses only on the corrections generated by GPT-3.5 in the zero-shot setting. These corrections are evaluated manually, with LMs, and with the Scribendy Score. This section also examines the semantic similarity between the corrections generated by GPT-3.5 and the original sentences. Finally, Section 6 summarizes the findings and outlines directions for future research.

2. Related Work

Grammatical error correction is usually approached as a monolingual sequence-to-sequence task: rewrite an incorrect sentence into its corrected version. This allows us to use models developed for machine translation (Ji et al., 2017; Junczys-Dowmunt et al., 2018; Chollampatt and Ng, 2018; Yuan et al., 2019; Grundkiewicz et al., 2019; Zhao et al., 2019; Kiyono et al., 2019; Choe et al., 2019; Kaneko et al., 2020; Rothe et al., 2021). As the method requires a large amount of data, models are commonly pre-trained on synthetic datasets. There are various techniques for augmenting error-free data, e.g., by adding random word perturba-

tions, Wikipedia edits, inserting errors by back translation, or imitating real error patterns (Kiyono et al., 2019; Grundkiewicz et al., 2019; Choe et al., 2019; Takahashi et al., 2020; Li and He, 2021). Another approach involves tagging sentences with token transformation labels and subsequently editing them based on the predicted transformations (Malmi et al., 2019; Stahlberg and Kumar, 2020; Omelianchuk et al., 2020; Tarnavskiy et al., 2022). More work on the GEC is reviewed in (Bryant et al., 2022).

GEC can be approached using Generative Adversarial Networks—GANs (Raheja and Alikaniotis, 2020), where the neural MT-based generator fixes errors, while the discriminator judges the quality of corrections. Parnow et al. (2021) presented a GAN-like architecture that consists of an error sequence labeler as a generator and an error detector as a discriminator.

Korotkova et al. (2019) investigated zero-shot GEC using Transformer-based MT models. Fang et al. (2023) assessed ChatGPT’s performance for Chinese, English, and German GEC on the sentence level, and for English on the document level. ChatGPT showed the highest scores on fluency and ability to generate correct results beyond what is indicated by the reference in comparison with other GEC models, such as GEC-ToR (Omelianchuk et al., 2020), T5-based models (Rothe et al., 2021), and the commercial system Grammarly. Coyne and Sakaguchi (2023) tested GPT-3.5 and GPT-4 performance on English GEC across zero-shot and few-shot settings and showed the tendency of the models to over-edit, changing the meaning of the sentence during correction.

3. Experiments

3.1. Data

We use the official datasets for all languages. For English, we use ConLL-14 (Ng et al., 2014), BEA-2019 (Bryant et al., 2019), and JGLEG (Napoles et al., 2017) test sets only for evaluation. The datasets FCE (Yannakoudakis et al., 2011) and W&I (Bryant et al., 2019) were combined for fine-tuning. We use the test sets from AKCES-GEC (Náplava and Straka, 2019) for Czech, COWS-L2H (Yamada et al., 2020) for Spanish, and UA-GEC (Syvokon et al., 2023) for Ukrainian for evaluation. We use Falco-MERLIN (Boyd et al., 2014) and RULEC-GEC (Rozovskaya and Roth, 2019) for German and Russian, respectively: the training sets were used for fine-tuning, and test sets for evaluation. For Swedish, we use SweLL-gold corpus (Volodina et al., 2019), which contains 502 essays written by adult learners of Swedish at various proficiency levels. The essays are pseudo-

	Dataset	Train	Dev	Test
CZ	AKCES-GEC	42 210	2 485	2 676
DE	Falko-MERLIN	19 237	2 503	2 337
EN	BEA-19	-	-	4 477
EN	CoNLL-14	-	-	1 312
EN	JGLEG	-	-	747
EN	FCE, W&I	59 941	-	-
ES	COWS-L2H	6 159	1 230	822
RU	RULEC-GEC	4 980	2 500	5 000
SV	SweLL-gold*	-	-	8 553
UA	UA-GEC	31 046	-	2 704

Table 1: Datasets used for fine-tuning and evaluation. We sampled 60% of the sentences from shuffled SweLL-gold.

anonymized and manually annotated. Given that this corpus includes some personal information, it is under GDPR protection.⁴ For evaluation, we shuffle the sentences in the entire corpus and randomly sample 60% of them. Statistics on the datasets are provided in Table 1.

3.2. Settings

Zero-shot GEC: For each input sentence, we prompted the model to follow the minimum-edit principle for correcting erroneous sentences. GPT-3.5 tends to provide feedback on the input’s correctness or to comment on the corrections. Therefore, we followed the suggestion of Fang et al. (2023) and added input and output tags to format the model’s output.

We experimented with two types of prompts. The first type is GEC instruction:

Prompt 1: “Provide a grammatical correction for the following sentence indicated by `<input> ERROR </input>` tag, making only necessary changes. If the input text is already correct, return it unchanged. Output the corrected version directly without any comments and explanations. Remember to format your corrected output with the tag `<output> Your Corrected Version </output>`. Please start: `<input> ERROR </input>`”

The second prompt is a chain-of-thought (Kojima et al., 2022) prompt suggested by Fang et al. (2023). It requires the model to follow particular steps of input comprehension:

Prompt 2: “Please identify and correct any grammatical errors in the following sentence indicated by `<input> ERROR </input>` tag, you need to comprehend the sentence as a whole before identifying and correcting

⁴<https://gdpr-info.eu/>

any errors step by step while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Remember to format your corrected output results with the tag `<output> Your Corrected Version </output>`. Please start: `<input> ERROR </input>`”

In both cases, we provide the model with information about its role:

“You are a LANGUAGE⁵ grammatical error correction tool that can identify and correct grammatical errors in a text.”

We also require the model to *not* translate input into English in case it is in other languages: “Do not translate input sentences into English.”

For all languages, we utilize English prompts for consistency. Our objective is to standardize prompts to reduce variation in the input, focusing solely on the target language, thereby ensuring more controlled outcomes.

Fine-tuning GPT-3.5: Fine-tuning was performed for English, German, and Russian. We use the GPT-3.5-turbo model for all three languages, without altering any parameters.⁶ Training examples were presented in a conversational chat format, with the same system role as for the zero-shot setting. We experiment with using (1) a short prompt message of “Improve grammar in the following sentence: ERROR”⁷ and (2) a longer message, which is the same as Prompt 2 for the zero-shot setting.

Re-ranking with GPT-3.5: For re-ranking, we use pre-existing GEC models: a publicly available encoder-decoder model⁸ trained on synthetic data for English (Kiyono et al., 2019); T5-large model,⁹ which we pre-train on synthetic data and then fine-tune on the RULEC-GEC training set for Russian, following Rothe et al. (2021). Synthetic data was generated from WMT News Crawl monolingual training data (Bojar et al., 2017) using Aspell confusion sets, following the approach of Grundkiewicz et al. (2019). We generated 10M sentences, with

⁵Variable for the target language.

⁶<https://platform.openai.com/docs/guides/fine-tuning>

⁷For Russian and German: “Improve grammar in the following sentence, do not translate into English: ERROR”. ERROR is a variable to replace with an input sentence.

⁸<https://github.com/butsugiri/gec-pseudodata>

⁹<https://huggingface.co/ai-forever/ruT5-large>

Lang.	Dataset	Model	P	R	$F_{0.5}$	σ
CZ	AKCES-GEC	ChatGPT (zero-shot)	68.5	64.1	67.6	1.9
		Transformer (Náplava and Straka, 2019)	84.2	66.7	80.0	-
DE	Falco-MERLIN	ChatGPT (zero-shot) (Fang et al., 2023)	59.9	63.9	60.7	-
		ChatGPT (zero-shot)	63.5	63.2	63.4	0.7
		Transformer (Náplava and Straka, 2019)	78.2	59.9	73.7	-
		T5 xxl (Rothe et al., 2021)	-	-	76.0	-
EN	BEA-19 (test)	ChatGPT (zero-shot) (Fang et al., 2023)	32.1	70.5	36.1	-
		ChatGPT (zero-shot)	47.4	64.6	50.0	3.5
		ChatGPT (rerank)	66.2	61.5	65.2	-
		ESC (Qorib et al., 2022)	86.7	60.9	79.9	-
	CoNLL-14	ChatGPT (zero-shot) (Fang et al., 2023)	50.2	59.0	51.7	-
		ChatGPT (zero-shot)	55.8	58.5	56.3	0.7
		ChatGPT (rerank)	59.9	52.4	58.3	-
		ESC (Qorib et al., 2022)	81.5	43.8	69.5	-
ES	COWS-L2H	mT5	29.0	10.8	21.7	-
		ChatGPT (zero-shot)	28.1	40.8	29.9	2.2
RU	RULEC-GEC	ChatGPT (zero-shot)	29.2	51.5	32.0	3.0
		ChatGPT (rerank)	39.5	43.7	40.3	-
		T5 xxl (Rothe et al., 2021)	-	-	51.6	-
		ruGPT large + re-ranking (Sorokin, 2022)	73.3	27.3	55.0	-
		RuT5 large + re-ranking (Katinskaia and Yangarber, 2023)	-	-	68.2	-
SV	SweLL-gold*	ChatGPT (zero-shot)	60.1	49.4	57.6	1.8
UA	UA-GEC	ChatGPT (zero-shot)	25.8	36.2	27.4	0.6
		mT5 large (Syvokon and Romanyshyn, 2023)	76.8	61.4	73.1	-

Table 2: Performance measured with ERRANT for the BEA-19 dataset, and with the M2 Scorer for other datasets. Best scores across a language (across datasets for English) are in **bold**. State-of-the-art models are provided for comparison. Our results are underlined. All ChatGPT models in the table are GPT-3.5 versions. σ denotes standard deviation for $F_{0.5}$ measure across 3 requests.

the same parameters as in (Náplava and Straka, 2019). The T5 model was pre-trained until convergence with the following parameters: 3 GPU V100, pre-training for 1.48M steps with batch size 6, weight decay 0, learning rate $5e-5$.

We generate 5 hypotheses for each source sentence using beam search decoding. The list of hypotheses for each sentence was provided to GPT-3.5 together with a prompt.

Prompt 3: “Re-rank the following sentences according to their grammatical correctness: LIST.¹⁰ Do not explain your choice. Numerate re-ranked sentences following the format:

1. Sentence A
2. Sentence B
- etc.

Do not provide any comments, notes, or translations. Do not change the input sentences.”

4. Evaluation

Zero-shot: Given the variability of GPT-3.5’s output for the same input on each request, we av-

¹⁰A variable to replace with a list of hypotheses.

Annotator	P	R	$F_{0.5}$
A	67.2	58.8	65.3
B	65.3	59.2	64.0
C	69.6	68.5	69.4
D	61.9	70.0	63.4

Table 3: Performance of GPT-3.5 in zero-shot setting measured with ERRANT for BEA-19 test set, reported separately for all 4 reference annotations.

eraged the performance over three requests per input sentence, see ChatGPT (zero-shot) in Table 2. Prompt 2 showed better performance for all languages than Prompt 1. Thus, Table 2 shows performance with the Prompt 2. All $F_{0.5}$ scores for GPT-3.5 are higher than those of Fang et al. (2023) in zero-shot CoT settings, though their recall is higher for English and German. This difference is likely due to the fact that the model may have evolved in the meantime. For Spanish, we compared the performance of GPT-3.5 with the results presented in this GEC report.¹¹ We are not aware of any Swedish GEC model to compare with, whose performance was measured by ERRANT or the M2 Scorer.

For all languages, except Czech and Ukrainian,

¹¹https://github.com/abhisaary/spanish_gec

Lang.	Dataset	P	R	$F_{0.5}$
EN	CoNLL-14	45.6	41.4	44.7
	BEA-19 (test)	40.4	51.5	42.2
DE	Falco-MERLIN	54.5	42.9	52.5
RU	RULEC-GEC	43.5	17.2	33.3

Table 4: Results of fine-tuning GPT-3.5 for GEC. For BEA-19, the performance was measured using ERRANT, and with M2 Scorer for other datasets.

GPT-3.5 has higher recall than all other models, but lower precision. High recall indicates strong error-correction capabilities, but the proposed corrections may differ from the reference corrections used for automatic evaluation. For example, as shown in Table 3, GPT-3.5 performance on English varies when we run evaluations against the gold-standard references within the BEA-19 test set provided by different annotators. Such variability in performance highlights the limitations of reference-based evaluation: to obtain an accurate assessment of GEC performance, it is essential to have *multiple* gold-standard references for each input sentence across all examined languages.

As shown in Table 2, none of the GPT-3.5 models achieved the highest F -score across all languages, except for Spanish. This finding aligns with the results of Zhang et al. (2023), who evaluated the performance of fine-tuned LLaMA on various writing-related tasks, including assessment of grammaticality. The authors demonstrated that a fully fine-tuned LLaMA model with 13B parameters achieves a comparable performance to that of lightweight RoBERTa and T5 models. The latter models are much faster and cheaper both in the training and the inference phases.

Fine-tuning: Among the two tested prompts, the short prompt achieved higher performance, see results in Table 4. The performance of the fine-tuned models is significantly lower than other settings for all tested languages. Manual inspection of the suggested corrections revealed that numerous sentences in the test remain unchanged. Given that fine-tuning aims to instruct the model on the expected corrections (i.e., making minimal changes without altering the sentence’s meaning), we hypothesize that fine-tuning for GEC requires more careful preparation of training data and prompt selection. Notably, the results for Russian exhibit very low recall, which could be due to the relatively small size of the training set and the presence of some number of uncorrected errors in the data.

Re-ranking: Re-ranking performance depends on the performance of the GEC models that are used to generate the hypotheses. In Table 5, we report the performance of English and Russian GEC models, mentioned in Subsection 3.2. Perfor-

Lang.	Dataset	Re-rank	P	R	$F_{0.5}$
EN	BEA-19 (test)	✗	66.2	61.5	65.2
	BEA-19 (test)	✓	47.4	64.6	50.0
	CoNLL-14	✗	68.6	44.9	62.1
	CoNLL-14	✓	59.9	52.4	58.3
RU	RULEC-GEC	✗	62.5	27.2	49.6
	RULEC-GEC	✓	38.4	42.5	39.1

Table 5: Re-ranking hypotheses with GPT-3.5. For BEA-19, performance was measured with ERRANT, and with M2 Scorer for other datasets.

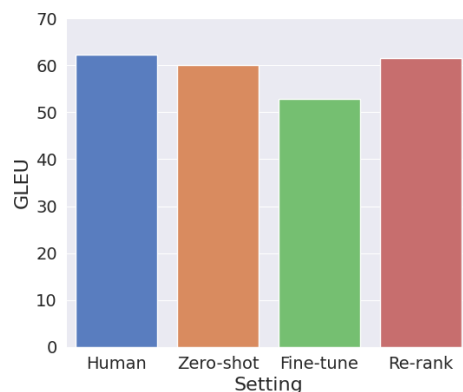


Figure 1: Performance for JFLEG, measured in GLEU.

mance before re-ranking is marked with ✗; performance after re-ranking—with ✓. As we can see, re-ranking with GPT-3.5 increases only recall, while decreasing precision. A possible explanation for this is that the model ranks the more fluent hypotheses with fewer errors higher, even if they deviate from the source sentences.

Table 2 shows that re-ranking correction hypotheses with GPT-3.5 performs better than in the zero-shot setting for GEC for English and Russian. Evaluation on JFLEG, a corpus for evaluating GEC systems with respect to fluency and grammatically, shows very high performance of GPT-3.5 as a re-ranker for English in comparison with a human-level evaluation score, see Figure 1. These observations indicate that GPT-3.5 may be more effectively used as a re-ranker in scenarios where other smaller fine-tuned GEC models are available and over-corrections/hallucinations are not tolerated.

5. Analysis of Zero-shot Output

Zero-shot (Manual evaluation): We carry out a preliminary manual evaluation for English and Russian. We randomly pick 200 sentences for each language: 70% of the sentences corrected by GPT-3.5, and 30% reference sentences. Two annotators, native speakers of both languages, were not informed about the origin of the evaluated sentences. Their task was to assess whether a sentence is

Original:	When I have spare time, I often gather my friends to watch basketball match on television.
Corrected:	When I have spare time, I often gather my friends to watch basketball match on television.
Reference:	When I have spare time, I often gather together my friends to watch basketball matches on television.
Original:	He is clearly aware of the dangers and brutalism of the sport, which is possibly why he enjoys it so much.
Corrected:	He is clearly aware of the dangers and brutalism of the sport, which is possibly why he enjoys it so much.
Reference:	He is clearly aware of the dangers and brutality of the sport, which is possibly why he enjoys it so much.
Original:	However, things do n't happened like what I hoped.
Corrected:	However, things don't happen like what I hoped.
Reference:	However, things didn't happen like I hoped.
Original:	I can speak English and I am a lovely, energetic and hardworking person besides I have recommended paper...
Corrected:	I can speak English, and I am a lovely and hardworking person. Besides, I have a recommended paper...
Reference:	I can speak English and I am a lovely and hardworking person. In addition, I have a reference...
Original:	которые образованы о обеспечивании государства высокого места в международной иерархии
Corrected:	которые разбираются в обеспечении государства высокого места в международной иерархии
Reference:	которые образованы в области обеспечения государству высокого места в международной иерархии (<i>who are really educated in the field of providing the state with a high place in the international hierarchy</i>)
Original:	Она всегда в модной одежде и ведет себя властью.
Corrected:	Она всегда в модной одежде и ведет себя с властью.
Reference:	Она всегда в модной одежде и ведет себя властно. (<i>She is always in fashionable clothes and acts bossy.</i>)
Original:	Я хочу привести яркий пример, о том, что может случиться...
Corrected:	Я хочу привести яркий пример о том , что может случиться...
Reference:	Я хочу привести яркий пример того , что может случиться... (<i>I want to give a vivid example of what could happen...</i>)
Original:	В курсе Основы светской этики учатся долг и совесть, принципы морали..
Corrected:	В курсе "Основы светской этики" учатся долгу и совести, принципы морали..
Reference:	В курсе "Основы светской этики" изучаются долг и совесть, принципы морали... (<i>The course "Fundamentals of secular ethics" covers duty and conscience, the principles of morality...</i>)

Table 6: Examples of erroneous corrections (see **Corrected**) which were suggested by GPT-3.5 in a zero-shot setting for English and Russian. Red denotes errors, blue denotes corrections suggested by the annotators.

grammatically acceptable.

The evaluation shows that 86.5% of the English sentences and 80% of the Russian sentences are acceptable for the annotators. For English, the majority of the errors are related to usage of hyphen, preposition, article, and tense. For Russian, most errors relate to punctuation, tense, and lexical compatibility. The model struggles to correct errors in constructions with government relations: errors in the usage of case of nouns and adjectives governed by verbs, or in prepositions required by particular verbs.

Table 6 shows examples of erroneous corrections proposed by GPT-3.5 and references suggested by the annotators. Interestingly, although GPT-3.5 tends to propose fluent corrections, the examples demonstrate an absence of agreement between proposed corrections and the semantics of the whole sentence.

Results of manual evaluation agree with error-type performance evaluated using ERRANT, see detailed error performance by error types for English (Table 7) and Russian (Table 8). Red denotes error types that were noted as problematic by the annotators.

Evaluation with LM: For automatic reference-free evaluation, we explore the use of pre-trained

LMs. The use of pre-trained LMs for assessing grammaticality has been proposed by several researchers (Linzen et al., 2016; Goldberg, 2019). Prior work has described incorporating LMs as components into GEC systems to score corrections, or to filter out low-quality examples from data (Lichtarge et al., 2020; Yasunaga et al., 2021). The core approach described in (Yasunaga et al., 2021) involves using an LM critic to obtain realistic pairs of grammatical and erroneous sentences from unlabeled data. The role of the critic is to assess whether an input sentence is correct by utilizing the LM's probability score. We adopt Yasunaga et al. (2021)'s intuition, that for a grammatical sentence, x_{good} , and its ungrammatical version with the same intended meaning, x_{bad} , we expect $p(x_{bad}) < p(x_{good})$, where p denotes the LM's probability normalized by the sentence length.

As the pre-trained LMs, we use GPT-2 models for Czech, German, Spanish, and Ukrainian. We use GPT-2 xl for English, GPT-3 large for Russian, and GPT-Sw3 with 6.7B parameters for Swedish. All models are openly available on HuggingFace. To evaluate GPT-3.5's quality, we aimed to use other models, as relying on GPT-3.5's own probabilities for its evaluation would be self-referential. We chose monolingual models for this purpose, given their exposure only to language-specific data.

Error type	<i>P</i>	<i>R</i>	<i>F</i> _{0.5}
ADJ	48.8	46.5	48.3
ADJ:FORM	71.4	76.9	72.5
ADV	43.9	51.3	45.3
CONJ	43.6	53.1	45.2
CONTR	92.9	86.7	91.6
DET	69.1	70.9	69.5
MORPH	75.7	72.8	75.1
NOUN	36.6	37.5	36.8
NOUN:INFL	84.2	84.2	84.2
NOUN:NUM	68.7	83.2	71.2
NOUN:POSS	78.4	52.7	71.4
ORTH	86.3	62.6	80.2
OTHER	43.9	31.9	40.8
PART	68.6	82.7	71.0
PREP	64.6	67.2	65.2
PRON	65.0	59.8	63.9
PUNCT	72.3	59.6	69.3
SPELL	79.1	78.2	78.9
VERB	53.0	42.3	50.5
VERB:FORM	73.7	81.6	75.1
VERB:INFL	100.0	87.5	97.2
VERB:SV	80.0	90.0	81.8
VERB:TENSE	59.4	58.8	59.3
WO	65.2	48.9	61.1

Table 7: Error-type performance of GPT-3.5 in zero-shot setting for BEA-19 test set (English), measured using ERRANT.

First, we collected pairs of erroneous source sentences (*Src*) and their gold-standard references (*Trg*) from multiple datasets. We verified that the pre-trained LMs assign higher probabilities to the gold-standard references than the source sentences; see Table 9, column $p(Trg) > p(Src)$ shows the percentage of instances in the dataset where *Trg* LM’s probability is higher than *Src* LM’s probability. As we observe, grammaticality correlates with the LM’s probabilities across all languages: LM’s probabilities are lower for erroneous source sentences.

Next, we compare probabilities assigned to GPT-3.5’s corrections (*Hyp*) with the probabilities of the gold-standard references; see column $p(Hyp) > p(Trg)$ in Table 9. GPT-3.5 suggestions received higher LM’s probabilities than the gold references only for English (over 54%) and Russian (50.2%). For the other five languages, the GPT-3.5 corrections have lower LM’s probabilities, which means a potentially lower level of grammaticality.

Evaluation with Scribendy Score: Another reference-less metric that we used is the Scribendy score (Islam and Magnani, 2021). It incorporates perplexity scores, a token sort ratio, and the Levenshtein distance score, and has been found to strongly correlate with human rankings. Since the Scribendy score is language-independent, we can use it for each of the tested languages.

Perplexity indicates whether the suggested change improves the grammaticality and fluency

Error type	<i>P</i>	<i>R</i>	<i>F</i> _{0.5}
ADJ	8.5	23.4	9.7
ADJ:CASE	20.2	59.1	23.3
ADJ:COMP_FORM	20.0	25.0	20.8
ADJ:FULL/SHORT	32.3	56.6	35.3
ADJ:GEN	40.8	62.5	43.9
ADJ:INFL	14.3	60.0	16.9
ADJ:NUM	30.0	58.1	33.2
ADV	11.8	25.4	13.3
CCONJ	6.8	30.6	7.9
MORPH	43.1	63.5	46.1
NOUN	16.5	28.2	17.9
NOUN:CASE	37.2	69.5	41.0
NOUN:INFL	66.7	80.0	68.9
NOUN:NUM	28.7	53.9	31.6
NUMERAL	14.6	63.7	17.2
NUMERAL:CASE	66.7	40.0	58.8
NUMERAL:INFL	75.0	100.0	78.9
ORTH	11.3	68.9	13.5
OTHER	5.3	17.5	6.1
PREP	22.3	41.5	24.6
PRON	16.0	32.8	17.8
PUNCT	14.6	58.0	17.2
SPELL	73.5	78.8	74.5
VERB	15.6	35.1	17.6
VERB:ASPECT	22.2	27.2	23.0
VERB:FORM	40.0	66.7	43.5
VERB:GEN	45.0	69.2	48.4
VERB:INFL	48.6	90.0	53.6
VERB:NUM	46.4	65.3	49.2
VERB:TENSE	12.8	27.8	14.4
VERB:VOICE	100.0	50.0	83.3
WO	4.8	54.6	5.8

Table 8: Error-type performance of GPT-3.5 in zero-shot setting for RULEC-GEC test set (Russian), measured using ERRANT.

of the sentence; it is calculated using the same LMs mentioned above. The token sort ratio and the Levenshtein distance between the source and the prediction are incorporated into the formula to ensure that the overall meaning of the sentence remains unchanged and to penalize predictions with repetitions.

The rightmost column of Table 9 displays the percentage of GPT-3.5’s suggestions that have higher Scribendy scores than the gold references. The results align with the evaluation using LMs: for English and Russian slightly over 50% of the hypotheses proposed by GPT-3.5 are ranked higher. As observed, there is a correlation between the two reference-less metrics, although the Scribendy score penalizes predictions that significantly rephrase the source sentences.

Semantic Similarity: For additional automatic evaluation of the changes proposed by GPT-3.5 during grammatical error correction, we compared the embeddings of the source sentences, the reference (or target) sentences provided by human annotators, and the corrections (or hypotheses) suggested by GPT-3.5. For this purpose, we uti-

Lang.	LM	Dataset	$p(Trg) > p(Src)$	$p(Hyp) > p(Trg)$	$S(Hyp) > S(Trg)$
CZ	GPT-2	AKCES-GEC	71.9	25.6	12.2
DE	GPT-2	Falco-MERLIN	91.9	33.6	18.4
EN	GPT-2 (xl)	BEA-19 (dev.)	90.3	54.7	52.5
		CoNLL-14	86.0	57.2	50.6
ES	GPT-2	COWS-L2H	90.0	41.3	42.1
RU	GPT-3	RULEC-GEC	90.5	50.2	50.1
SV	GPT-Sw3 6.7B	SweLL-gold*	93.1	30.7	27.4
UA	GPT-2	UA-GEC	68.3	35.3	26.2

Table 9: Evaluation: how probabilities assigned to a sentence by pre-trained LM correlate with sentence type: *Src* refers to original erroneous sentence; *Trg* refers to gold-standard reference; *Hyp* refers to correction suggested by GPT-3.5 in zero-shot setting. Reported numbers are percentages. *S* in last column denotes Scribendy score.

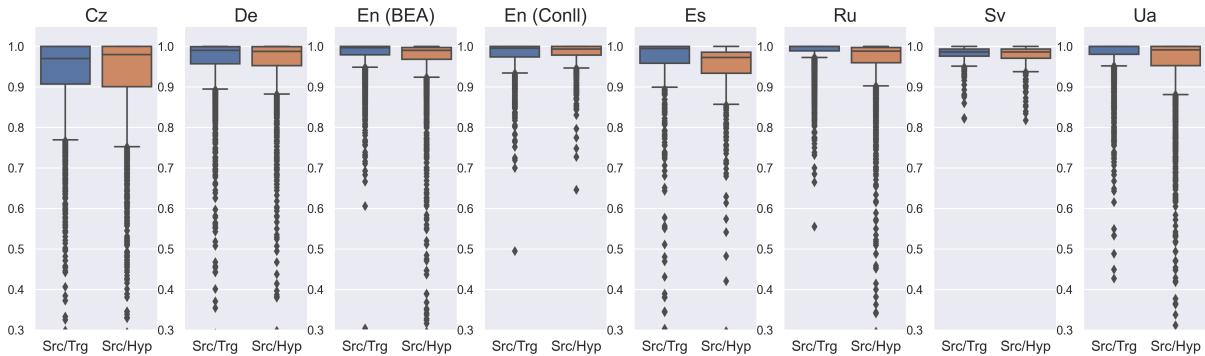


Figure 2: Cosine similarities between the source sentence embeddings and the target sentence embeddings (*Src/Trg*) and cosine similarities between the source sentence embeddings and sentence embeddings of the hypotheses suggested by GPT-3.5 (*Src/Hyp*).

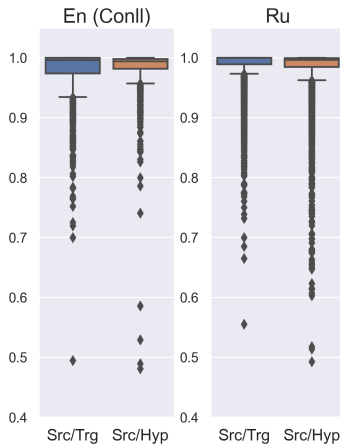


Figure 3: Cosine similarities between source sentence embeddings and target sentence embeddings (*Src/Trg*) and between the source sentence embeddings and sentence embeddings of the re-ranked hypotheses suggested by other GEC models (*Src/Hyp*).

lize the Sentence Transformer model,¹² a modification of the pretrained BERT that uses siamese

¹²<https://www.sbert.net/>

and triplet network structures to derive semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). For all languages, we use a multilingual model¹³ to encode sentences (Reimers and Gurevych, 2020). After that, we measured the cosine similarity (1) between the source embeddings and the target embeddings (*Src/Trg*), and (2) between the source embeddings and the hypotheses embeddings (*Src/Hyp*), to assess the semantic similarity of GPT-3.5’s suggestions to the original input sentences.

From Figure 2, we observe that the difference between two cosine similarities (*Src/Trg* and *Src/Hyp*) varies across languages. In the case of Czech and German, both the gold references and GPT-3.5 correction hypotheses exhibit significant differences from the source sentences. This similarity in the deviations from the source sentences, observed in both the gold references and the model’s suggestions, implies extensive corrections made by humans and by the model. Although GPT-3.5 is known for occasionally altering the meaning of the source sentences, its corrections for English can be semantically even closer to the source sentence than the gold references, e.g., see the results for

¹³distiluse-base-multilingual-cased-v2

CoNLL. For Russian, Spanish, and Ukrainian, the plots indicate that changes proposed by the model tend to transform the sentence meaning the most.

Compare the cosine similarities for CoNLL and for Russian RULEC in Figure 2 with the analogous plots in Figure 3, where *Hyp* denotes corrections generated by two other GEC models and re-ranked by GPT-3.5. We can see that the semantics of the corrected sentences remain close to the semantics of the source sentences. This is consistent with the observation that selecting top-ranked hypotheses results in higher precision, despite yielding a lower recall when compared to a zero-shot setting (see *rerank* and *zero-shot* in Table 2).

Cosine similarities between sentence embeddings, combined with LM scores and the Scribendy scores, as well as evaluation of GEC performance using reference-based metrics, suggests that GPT-3.5 may not be the optimal GEC model for Czech, German, Spanish, and Ukrainian. For these languages, the overall GEC performance of GPT-3.5 is lower: LMs scores show that only a small percentage of suggested corrections are less “surprising” to language models than the gold reference sentences; the Scribendy scores are also modest, which means that sentences lexically deviate from the input and might have repetitions.

For Russian, the results indicate that GPT-3.5 can correct more errors than any other model (see the highest recall for RULEC in Table 2), but this comes at the expense of reduced semantic similarity between the source sentences and the proposed corrections. Performance for English is the most encouraging: although precision is lower compared to other GEC models, recall is higher, and the semantic similarity between the source sentences and suggested corrections is notably high. Re-ranking achieves even higher similarity while maintaining high recall, since GPT-3.5 prioritizes fluency. Performance for Swedish is also promising, though we cannot compare it with other GEC models. Corrections suggested by GPT-3.5 stay semantically similar to the source sentences, which may also result from under-corrected errors—this will agree with lower LM and Scribendy scores assigned to the correction hypotheses.

These findings suggest that while GPT-3.5 can be used for GEC, its efficacy varies across languages. For certain languages, it might offer corrections that alter the intended semantics of the sentence.

6. Conclusions

In this work, we evaluate the performance of GPT-3.5 models for GEC across different languages and experimental settings. To our knowledge, this is the first paper that evaluates GPT-3.5 for GEC across

the number of languages that we cover. The results indicate that GPT-3.5 has strong error-correction capabilities for various languages. However, current reference-based evaluation methods tend to underestimate its performance, because GPT-3.5 generates fluent sentences, often not following the minimum-edit principle. We perform several types of evaluation of the corrections suggested by GPT-3.5: manually and by using reference-less metrics focused on sentence fluency and semantic similarity to the original input.

We find that GPT-3.5’s behavior varies by language. For Czech, German, Russian, Spanish, and Ukrainian, it tends to modify the input more freely, not always achieving superior performance. In contrast, for English, GPT-3.5 achieves high recall, generates fluent sentences, and does not significantly alter sentence semantics. For Swedish, GPT-3.5 archives relatively high precision and generates corrections that do not deviate semantically from the input sentences, which may, though, result from under-corrected errors, therefore, it requires manual evaluation.

The results of our experiments show that using GPT-3.5 to re-rank correction hypotheses generated by other GEC models outperforms GPT-3.5’s own performance in both the zero-shot and fine-tuning settings for English and Russian. This indicates that GPT-3.5 can be employed in a setup with smaller fine-tuned GEC models. Manual evaluation of the corrections suggested by GPT-3.5 for English and Russian shows that the model often struggles with correcting: punctuation, articles, preposition errors, and verb tense errors in English; and errors in punctuation, in government between words, and in lexical compatibility in Russian.

Based on our evaluation, we emphasize the need to develop new evaluation metrics for GEC. Traditional reference-based metrics may not adequately capture the advancements of modern LLMs. We propose that a new metric should be based on multiple gold-standard references from human annotators, and it should take into account to what extent the *meaning* of the corrected sentence deviates from the meaning of the input.

Considering that GPT-3.5 can be further improved for various error types (particularly for languages other than English) and the challenges in evaluating its correctness, we find that it has the potential for integration into GEC systems, such as an ensemble with other models, serving as a re-ranker of correction hypotheses. In future work, we plan to broaden our studies to more languages, experiment with freely available large language models, and perform a thorough in-depth human evaluation for all languages under study.

Limitations

This work presents preliminary experiments on GEC for several languages using GPT-3.5 models, which were started by other researchers. Our experiments in a zero-shot setting are limited to using only two prompt messages for all languages, as experimenting with various prompts for different languages was not the focus of our work. Experiments with fine-tuning were done with only three languages, without changing any default parameters due to financial limitations. In the future, we plan to utilize freely-available large language models. As for re-ranking hypotheses generated by other GEC models, we had resources to train only Russian GEC model; English GEC model was available for public use. We plan to extend re-ranking experiments with the other languages in future work. We also did not experiment with providing GPT-3.5 with any correction examples, as was done in (Fang et al., 2023).

Our manual evaluation of generated corrections is limited: we were unable to involve more native annotators for English and Russian and to perform any manual evaluation for other languages. In future work, we hope to have more resources to invite more annotators. However, we find estimation of GEC performance problematic: reference-based methods do not allow to get a realistic evaluation, while manual evaluation is very expensive and time-consuming, especially if the number of languages is large.

We used only publicly available resources for all conducted experiments. Regarding GPT-3.5, despite the official API being available to any registered user, we found a number of difficulties in using it in research experiments: black-box models are constantly updated, making it hard to verify any results obtained with them; it is difficult to control the model's behavior in a zero-shot setting because it frequently does not follow the required format; experiments with fine-tuning can be expensive and not affordable to many researchers.

Ethics Statement

We use only publicly available resources for all conducted experiments. All annotators were volunteer students who performed the tasks as a part of their studies and received credits for it.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia,

and Marco Turchi. 2017. [Findings of the 2017 Conference on Machine Translation \(WMT17\)](#). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Bryant, Mariano Felice, 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. 2022. Grammatical error correction: A survey of the state of the art. *arXiv preprint arXiv:2211.05166*.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeol Yoon. 2019. [A neural grammatical error correction system built on better pre-training and sequential transfer learning](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of GPT-3's performance in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

- 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 preprint arXiv:2304.01746*.
- Yoav Goldberg. 2019. Assessing BERT’s syntactic abilities. *arXiv preprint arXiv:1901.05287*.
- Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. [Neural grammatical error correction systems with unsupervised pre-training on synthetic data](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.
- Md Asadul Islam and Enrico Magnani. 2021. [Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. [A nested attention neural hybrid model for grammatical error correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–762, Vancouver, Canada. Association for Computational Linguistics.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. [Approaching neural grammatical error correction as a low-resource machine translation task](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.
- Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. [Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.
- Anisia Katinskaia and Roman Yangarber. 2023. [Grammatical error correction for sentence-level assessment in language learning](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 488–502, Toronto, Canada. Association for Computational Linguistics.
- Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. [An empirical study of incorporating pseudo data into grammatical error correction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Elizaveta Korotkova, Agnes Luhtaru, Maksym Del, Krista Liin, Daiga Deksnė, and Mark Fishel. 2019. Grammatical error correction and style transfer via zero-shot monolingual translation. *arXiv preprint arXiv:1903.11283*.
- Xia Li and Junyi He. 2021. [Data augmentation of incorporating real error patterns and linguistic knowledge for grammatical error correction](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 223–233, Online. Association for Computational Linguistics.
- Jared Lichtarge, Chris Alberti, and Shankar Kumar. 2020. [Data weighted training strategies for grammatical error correction](#). *Transactions of the Association for Computational Linguistics*, 8:634–646.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the Ability of LSTMs to Learn Syntax-sensitive Dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- 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.

- Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. [Encode, tag, realize: High-precision text editing](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.
- Jakub Náplava and Milan Straka. 2019. [Grammatical error correction in low-resource scenarios](#). In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.
- Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. [JFLEG: A fluency corpus and benchmark for grammatical error correction](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. [The CoNLL-2014 shared task on grammatical error correction](#). In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.
- Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhashnyi. 2020. [GECToR – grammatical error correction: Tag, not rewrite](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv:2303.08774*.
- Kevin Parnow, Zuchao Li, and Hai Zhao. 2021. [Grammatical error correction as GAN-like sequence labeling](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3284–3290, Online. Association for Computational Linguistics.
- Muhammad Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. [Frustratingly easy system combination for grammatical error correction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.
- Vipul Raheja and Dimitris Alikaniotis. 2020. [Adversarial Grammatical Error Correction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3075–3087, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. [A simple recipe for multilingual grammatical error correction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.
- 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.
- Alexey Sorokin. 2022. [Improved grammatical error correction by ranking elementary edits](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11416–11429, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Felix Stahlberg and Shankar Kumar. 2020. [Seq2Edits: Sequence transduction using span-level edit operations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online. Association for Computational Linguistics.

- Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmichuk, 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.
- Oleksiy Syvokon and Mariana Romanyshyn. 2023. The UNLP 2023 shared task on grammatical error correction for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 132–137.
- Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. [Grammatical error correction using pseudo learner corpus considering learner’s error tendency](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, Online. Association for Computational Linguistics.
- Maksym Tarnavskiy, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. [Ensembling and knowledge distilling of large sequence taggers for grammatical error correction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852, Dublin, Ireland. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: open and efficient foundation language models.
- Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. 2019. The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.
- Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.
- Aaron Yamada, Sam Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae, and Claudia Sánchez-Gutiérrez. 2020. COWS-L2H: A corpus of Spanish learner writing. *Research in Corpus Linguistics*, 8(1):17–32.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. [LM-Critic: Language models for unsupervised grammatical error correction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zheng Yuan, Felix Stahlberg, Marek Rei, Bill Byrne, and Helen Yannakoudakis. 2019. [Neural and FST-based approaches to grammatical error correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 228–239, Florence, Italy. Association for Computational Linguistics.
- Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv preprint arXiv:2305.13225*.
- Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. [Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.