Evaluating GPT-3.5 and GPT-4 on Grammatical Error Correction for Brazilian Portuguese

Maria Carolina Penteado [*] **Fábio Perez** [*]

Abstract

We investigate the effectiveness of GPT-3.5 and GPT-4, two large language models, as Grammatical Error Correction (GEC) tools for Brazilian Portuguese and compare their performance against Microsoft Word and Google Docs. We introduce a GEC dataset for Brazilian Portuguese with four categories: Grammar, Spelling, Internet, and Fast typing. Our results show that while GPT-4 has higher recall than other methods, LLMs tend to have lower precision, leading to overcorrection. This study demonstrates the potential of LLMs as practical GEC tools for Brazilian Portuguese and encourages further exploration of LLMs for non-English languages and other educational settings.

1. Introduction

Large language models (LLMs) have revolutionized the field of natural language processing by enabling computers to process and generate human-like language (Kasneci et al., 2023). LLMs have the potential to be particularly useful for Grammatical Error Correction (GEC) (Wu et al., 2023; Bryant et al., 2022) and can be a valuable educational tool to enhance students' writing skills by providing real-time feedback and corrections.

Traditional GEC methods usually rely on pre-defined rules to identify and correct errors. While these methods can effectively detect simple misspellings, they may struggle to correct more complex grammatical errors. In contrast, LLMs can model language from large amounts of text data, which could lead to more natural and contextually appropriate corrections. By analyzing the context and meaning of a sentence, LLMs may identify errors that traditional methods may miss and provide more nuanced corrections.

Although large language models (LLMs) have gained widespread attention for their performance in English language applications, recent studies have shown that they can produce good results for other languages. While the amount of data available for training LLMs in languages other than English is often more limited, the success of these models in tasks such as translation, language modeling, and sentiment analysis demonstrates their potential for improving language processing across a range of different languages.

In this work, we take an initial step on investigating the effectiveness of GPT-3.5 and GPT-4 (OpenAI, 2023), two LLMs created by OpenAI, as a GEC tool for Brazilian Portuguese. Our main contributions are the following:

1. We compare GPT-3.5 and GPT-4 against Microsoft Word and Google Docs and show that LLMs can be a powerful tool for GEC.
2. We crafted a GEC dataset for Brazilian Portuguese, including four categories: *Grammar*, *Spelling*, *Internet*, and *Fast typing*.
3. We quantitative and qualitatively evaluated LLMs as a GEC tool for Brazilian Portuguese.

2. Related Work

Nunes et al. (2023) explored the use of GPT-3.5 and GPT-4 to answer questions for the *Exame Nacional do Ensino Médio* (ENEM), an entrance examination used by many Brazilian universities. They tested different prompt strategies, including using Chain-of-Thought (CoT) to generate explanations for answers, and found that GPT-4 with CoT was the best-performing approach, achieving an accuracy of 87% on the 2022 exam.

Wu et al. (2023) evaluated the performance of different models for GEC, including Grammarly, GECToR, and ChatGPT (authors did not specify whether they used GPT-3.5 or GPT-4), and found that automatic evaluation methods result in worse numbers for ChatGPT than other GEC methods. In contrast, human evaluation showed that ChatGPT produces fewer under-corrections or miscorrections and more over-corrections, indicating not only the potential of LLMs for GEC but also the limitation of automatic metrics to evaluate GEC tools.

[*]Equal contribution . Correspondence to: Fábio Perez <fabiovmp@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Fang et al. (2023) investigated GPT-3.5's potential for GEC using zero- and few-shot chain-of-thought settings. The model was evaluated in English, German, and Chinese, showcasing its multilingual capabilities. The study found that GPT-3.5 exhibited strong error detection and generated fluent sentences but led to over-corrections.

Despite their outstanding performance on many tasks, LLMs may not be the silver bullet for NLP in multi-lingual settings. Lai et al. (2023) evaluated ChatGPT on various NLP tasks and languages, showing that it performs significantly worse than state-of-the-art supervised models for most tasks in different languages, including English. Their work does not include GEC, and Portuguese is only evaluated for relation extraction.

The shortage of academic research on LLMs for multilingual settings, especially for Brazilian Portuguese, highlights the need for further engagement in this field. Our work aims to fill this gap by exploring the potential of GPT-3.5 and GPT-4 as GEC tools for Brazilian Portuguese.

3. Methodology

3.1. Dataset

We created the dataset (Table 1) by having native Brazilian Portuguese speakers manually write multiple sentences and dividing them into four categories: grammar, orthography, mistypes, and internet language. All categories list incorrect sentences and their correct pairs. The categories are described as follows:

- **Grammar** — 34 sets of three (totaling 102) phrases containing two words or expressions that are commonly swapped due to their similarity.
- **Spelling** — 100 sentences with spelling, punctuation, or accentuation errors.
- **Fast typing** — 40 mistyped (e.g., when typing too fast) sentences.
- **Internet language** — 40 sentences containing slang, abbreviations, and neologisms often used in virtual communication.

We find it important to acknowledge that the dataset may reflect the biases of the human curators and may not fully encompass the complexities and variations present in real-world data. However, the limited availability of corpora specifically designed for GEC in Brazilian Portuguese compelled us to create our dataset, which, despite its potential limitations, represents a starting point in the task.

The dataset is available in the supplementary material.

3.2. Experiments

We compared GPT-3.5 and GPT-4, two LLMs, against the spelling and grammar error correction features on Google Docs and Microsoft Word, two widely-used text editors.

For Google Docs (docs.google.com), we first set the language on *File* → *Language* → *Português (Brasil)*. Then we selected *Tools* → *Spelling and grammar* → *Spelling and grammar check*. Finally, for every error, we clicked on *Accept*.

We used the online version of Microsoft Word (onedrive.live.com). First, we set the language on *Set Proofing Language* → *Current Document* → *Portuguese (Brazil)*. Then, we opened the *Corrections* tab and selected all errors under *Spelling and Grammar*. For each error, we selected the first suggestion. We repeated the process until Word stopped finding errors.

For GPT-3.5 and GPT-4, we used ChatGPT (chat.openai.com) with the prompt shown in Table 2. We shuffled the phrases and ensured the same pair of correct and incorrect phrases did not appear in the same prompt. Instead of running phrases individually, we ran 20 to 26 simultaneous phrases in one prompt, depending on the category. We used the ChatGPT interface and not the OpenAI API since we did not have access to the GPT-4 API at the time of the experiments. We did not focus on optimizing the prompt as our goal is to evaluate the usefulness of LLMs for GEC in Brazilian Portuguese without requiring deep LLMs knowledge. We believe more careful prompt engineering may improve the results.

4. Results

CoNLL2014 (Ng et al., 2014) employs an evaluation method in which GEC tools are evaluated by all edits they made on phrases against gold-standard edits. Instead, we evaluate GEC tools by comparing the modified phrases against the gold-standard ones. For the *Grammar* and *Spelling* categories, we also ran GEC tools on phrases without grammatical errors to evaluate false positives. We calculated four metrics:

- **Precision** — From the phrases modified by the GEC tool, how many were successfully corrected?
- **Recall** — From the ungrammatical phrases, how many were successfully corrected by the GEC tool?
- $F_{0.5}$ **Score** — A metric that combines both precision and recall, but emphasizes precision twice as much as recall. It is commonly used in GEC studies (Ng et al., 2014).
- **True Negative Rate (TNR)** — From the grammatical phrases, how many were successfully not modified by the GEC tool?

Table 1. Description of the developed dataset, divided into four categories: *Grammar*, *Spelling*, *Fast typing*, and *Internet*. The table shows the total correct and incorrect phrases per category and example phrases from the dataset. *Grammar* and *Spelling* include only one error per phrase, while *Fast typing* and *Internet* include multiple.

	COR.	INC.	CORR. EXAMPLE	INCORR. EXAMPLE
GRAMMAR	102	102	Você nunca mais falou com agente	Você nunca mais falou com a gente
SPELLING	100	100	A análize do documento será feita por um advogado	A análise do documento será feita por um advogado
FAST TYPING	-	40	ele já quebru todos oc copos nvoos que comepri esse mêsd	Ele já quebrou todos os copos novos que comprei esse mês
INTERNET	-	40	n dá p escutar, n sei o pq	Não dá para escutar, não sei o porquê

Table 2. Prompt used for GPT-3.5 and GPT-4 and its English translation as reference. We prompted both models to add *[Correta]* if the phrase is correct to avoid them appending long texts saying the phrase is correct. We removed any *[Correta]* occurrence before evaluating the models.

PROMPT
Corrija os erros gramaticais das seguintes frases em Português brasileiro. Não altere o significado das frases, apenas as corrija. Não altere frases gramaticalmente corretas, apenas escreva [Correta] após a frase. {list of phrases}
PROMPT TRANSLATION TO ENGLISH
Fix the grammatical errors in the following Brazilian Portuguese sentences. Do not change the meaning of the sentences, just fix them. Do not change grammatically correct sentences, just write [Correct] after the sentence.

We evaluated *Grammar* and *Spelling* using the four metrics and *Internet* and *Fast typing* using recall. Table 3 shows the results for all experiments. We define true/false positive/negative as follows (see Table A1 for examples):

- **True Positive (TP)** — incorrect phrase is corrected by the GEC tool.
- **False Positive (FP)** — correct phrase is wrongly modified by the GEC tool.
- **True Negative (TN)** — correct phrase is not modified by the GEC tool.
- **False Negative (FN)** — incorrect phrase is not corrected by the GEC tool.

5. Discussion

Results (Table 3) for *Grammar* and *Spelling* show that GPT-3.5 and GPT-4 have superior recall and worse precision than Microsoft Word and Google Docs. These results agree with those by Wu et al. (2023) and Fang et al. (2023) and suggest that while GPT models are very effective at identifying errors, they tend to make more corrections than necessary, potentially altering the meaning or style of the text. The lower TNR values also confirms that LLMs tend to modify correct phrases.

One possible explanation for the higher recall of LLMs is their ability to model language from large amounts of text data, allowing them to capture a wide range of language patterns and contextual nuances. This makes them effective at detecting complex grammatical errors, but their open-ended nature can lead to overcorrection by generating multiple possible corrections without clearly picking the most appropriate one. Furthermore, LLMs may have lower precision because they often prioritize fluency and coherence over grammatical accuracy, leading to unnecessary changes to the text, increasing false positives. In contrast, rule-based methods prioritize grammatical accuracy and make changes only when necessary.

Although strongly impacted by the lower precision, GPT-4 shows a higher $F_{0.5}$ score than any other methods for both *Grammar* and *Spelling*. GPT-3.5, however, has lower $F_{0.5}$ scores than Google Docs and Microsoft Word, indicating that GPT-4 is a clear improvement over GPT-3.5 as a GEC tool for Brazilian Portuguese.

Finally, GPT-3.5 and GPT-4 perform much better than Microsoft Word and Google Docs for the *Internet* and *Fast typing* categories. Traditional methods struggle with these tasks as they are strongly context-dependent, while LLMs thrive due to being trained on vast amounts of text. This demonstrates the capabilities of LLMs as a GEC tool for non-traditional GEC scenarios.

Table 3. Evaluation results for all experiments. Since the results are not deterministic, values for GPT-3.5 and GPT-4 represent the average and standard deviation for three runs.

		MS WORD	GOOGLE DOCS	GPT-3.5	GPT-4
INTERNET	RECALL	12.5%	5.0%	78.3±1.3%	**89.3±1.3%**
FAST TYPING	RECALL	27.5%	40.0%	85.0±0.0%	**90.0±1.3%**
GRAMMAR	PRECISION	89.1%	**97.4%**	67.5±0.2%	86.8±0.7%
	RECALL	40.2%	36.3%	63.7±1.7%	**75.5±1.7%**
	$F_{0.5}$	71.7%	72.8%	66.7±0.5%	**84.3±1%**
	TNR	95.1%	**99.0%**	69.3±0.6%	88.5±0.6%
SPELLING	PRECISION	94.9%	**100%**	79.7±1.7%	99.3±0.6%
	RECALL	74.0%	66.0%	85±3.5%	**92.0±6.1%**
	$F_{0.5}$	89.8%	90.7%	80.7±2%	**97.7±1.8%**
	TNR	96.0%	**100%**	78.3±1.5%	99.3±0.6%

We also performed a qualitative analysis by checking each correction provided by GPT-3.5 and GPT-4. We identified four explicit behaviors. See Table A2 for examples of phrases for each behavior.

The first behavior (over-correction) considers extra edits that lead to grammatically correct sentences without meaning changes (e.g., add/remove commas, convert commas into semicolons, and upper vs. lower case). GPT-3.5 delivered 54 (out of 484) sentences with such behavior vs. six from GPT-4. The second behavior (omission) refers to models failing to detect errors and occurred 22 and 23 times on GPT-3.5 and GPT-4, respectively.

The third behavior (grammatical miscorrection) includes changes that adhere to grammatical rules but modify the sentence's meaning (e.g., removing/adding/substituting words and inverting the order of excerpts). GPT-3.5 corrections fell in this category 41 times vs. 13 times for GPT-4. Finally, the fourth behavior (ungrammatical miscorrection) is similar to the previous one but leads to ungrammatical sentences. GPT-3.5 and GPT-4 produced 3 and 1 outputs in this category, respectively.

5.1. Limitations and Challenges of LLMs as GEC tools

While large language models (LLMs) have shown considerable promise for Grammatical Error Correction (GEC), limitations and challenges must be considered when using these models for GEC tasks.

Open-endedness LLMs are open-ended and stochastic by nature. Unlike rule-based models, LLMs generate text based on patterns learned from training data. This can make it difficult to constrain the model, resulting in the replacement of grammatically correct words with other words that may occur more frequently in a given context (Bryant et al., 2022). Another unpredictability of LLMs is their tendency to produce "hallucinations" – outputs that are not necessarily true or based on the input data (OpenAI, 2023). This can result in the generation of incorrect or irrelevant corrections.

Prompt engineering LLMs' performance rely on the used prompts (Brown et al., 2020), where LLM-based GEC tools might need prompt engineering to achieve high-quality outputs. The effectiveness of a prompt may vary significantly depending on the task, and determining an optimal prompt may require extensive experimentation.

Hardware constraints The large-scale nature of LLMs requires powerful hardware, which can be a barrier for many users and institutions. This can make LLMs less accessible and cost-effective for GEC tasks, particularly for those with limited resources or budget constraints. To interact with LLMs that cannot run on consumer hardware, one must send requests to third-party servers, requiring an internet connection and posing a privacy risk.

Biases and malicious uses LLMs may contain biases and inaccuracies, posing a challenge in ensuring that corrections do not inadvertently perpetuate harmful stereotypes or misinformation (Blodgett et al., 2020; Nadeem et al., 2020; Garrido-Muñoz et al., 2021). LLMs may also suffer from malicious attacks intent to misleading the model (Perez & Ribeiro, 2022; Greshake et al., 2023).

6. Conclusion

Our study demonstrates the potential of LLMs as effective GEC tools for Brazilian Portuguese. We hope this work encourages further exploration of the impact of LLMs on Brazilian Portuguese and other non-English languages and spurs interest in developing and refining LLMs for diverse linguistic contexts. As a suggestion for future works, we believe that curating larger and better datasets that capture real-world data (e.g., by collecting grammatical errors made in real scenarios) could strengthen the field. Moreover, we encourage researchers to continue investigating the potential of LLMs in educational settings (see Appendix B).

References

Blodgett, S. L., Barocas, S., Daum'e, H., and Wallach, H. M. Language (technology) is power: A critical survey of "bias" in NLP. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., and Briscoe, T. Grammatical error correction: A survey of the state of the art. *arXiv preprint arXiv:2211.05166*, 2022.

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., and Zhang, Y. Is ChatGPT a highly fluent grammatical error correction system? a comprehensive evaluation, 2023.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A. A survey on bias in deep NLP. *Applied Sciences*, 11:3184, 2021.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *ArXiv*, abs/2302.12173, 2023.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, 2023.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14, 2014.

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. Evaluating GPT-3.5 and GPT-4 models on brazilian university admission exams. *arXiv preprint arXiv:2303.17003*, 2023.

OpenAI. GPT-4 technical report, 2023.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models, 2022.

Wu, H., Wang, W., Wan, Y., Jiao, W., and Lyu, M. ChatGPT or grammarly? evaluating ChatGPT on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*, 2023.

Table A1. Examples of TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative) results.

INPUT PHRASE	EXAMPLES OF RESULTS
Só encontrei ingressos para a última seção do filme. [INCORRECT]	[TP] Só encontrei ingressos para a última sessão do filme. [FN] Só encontrei ingressos para a última seção do filme. [FN] Só encontrei ingressos para a última seção daquele filme.
Só encontrei ingressos para a última sessão do filme. [CORRECT]	[TN] Só encontrei ingressos para a última sessão do filme. [FP] Só encontrei ingressos para a última seção do filme. [FP] Só encontrei ingressos para a última sessão daquele filme.
Mais será possível que ele tenha ido tão longe? [INCORRECT]	[TP] Mas será possível que ele tenha ido tão longe? [FN] Mais será possível que ele tenha ido tão longe? [FN] Mas será possível que ela tenha ido tão longe?
Mas será possível que ele tenha ido tão longe? [CORRECT]	[TN] Mas será possível que ele tenha ido tão longe? [FN] Mais será possível que ele tenha ido tão longe? [FN] Mas será possível que ela tenha ido tão longe?

A. Appendix – Example Tables

B. Appendix – More Educational Applications

In addition to their potential for being a powerful GEC tool for Brazilian Portuguese, LLMs hold promise for improving educational outcomes in Brazilian schools and universities. We list some of these ideas:

Better GEC Tools LLMs can not only correct ungrammatical phrases but also guide students in understanding why the phrases are ungrammatical and explaining how to fix them. See Table B3 for an example.

Hyper-personalized tutoring LLMs can facilitate hyper-personalized tutoring by adapting to each student's unique learning style, progress, and needs, providing tailored feedback and guidance to enhance the learning experience.

Learning disabilities LLMs can be used to optimize tools developed to assist students with learning disabilities, creating contents that would give them proper conditions to comprehend subjects, follow instructions and answer questions.

Classroom augmentation LLMs can be integrated into various tools and applications to support classroom teachers, learners, and test developers. For instance, they can be used to generate realistic and diverse practice questions or exercises and to assist in creating engaging and culturally-relevant educational content.

Grading assistant LLMs can assist in grading and analyzing students' written responses and provide insights into their understanding and misconceptions.

The potential of LLMs extends beyond English language applications, offering numerous possibilities for revolutionizing education in Brazilian Portuguese and other languages.

Table A2. Examples of identified behaviors in our qualitative analysis.

BEHAVIOR	CORRECT PHRASE	LLM RESULT
OVERCORRECTION	Quantas vezes vou precisar repetir até você me entender? As obras daquele diretor são impactantes, são filmes que nos fazem refletir de verdade. Podemos conversar, a fim de compreender melhor a questão.	Quantas vezes vou precisar repetir até você me entender [GPT-4] As obras daquele diretor são impactantes; são filmes que nos fazem refletir de verdade. [GPT-3.5] Podemos conversar a fim de compreender melhor a questão. [GPT-3.5]
OMISSION	Vou buscar meu relógio no conserto. Quanto ao recebimento do envelope, confirmei-o na mesma hora. Vamos construir um bom canal de comunicação, ao invés de continuar com essa dificuldade em estabelecer um diálogo.	Vou buscar meu relógio no concerto. [GPT-3.5] Quanto ao recebimento do envelope, confirmei-lhe na mesma hora. [GPT-4] Vamos construir um bom canal de comunicação, em vez de continuar com essa dificuldade em estabelecer um diálogo. [GPT-4]
GRAMMATICAL MISCORRECTION	Apesar de ela ter deixado todos os direcionamentos, não foi possível realizar a apresentação em sua ausência. O senhor encheu o balde de água. A astrologia classifica como Mercúrio retrógrado o período pelo qual estamos passando.	Apesar de ela ter deixado todas as orientações, não foi possível realizar a apresentação em sua ausência. [GPT-3.5] O senhor encheu o balde com água. [GPT-3.5] A astrologia classifica o período pelo qual estamos passando como Mercúrio retrógrado. [GPT-3.5]
UNGRAMMATICAL MISCORRECTION	Vou viajar nesse próximo final de semana. Buscaram, então, a melhor decisão para ambos os lados. As duas meninas, que foram tão amigas no passado, hoje já não têm mais contato.	Vou viajar neste próximo final de semana. [GPT-3.5][GPT-4] Buscaram então a melhor decisão para ambos os lados. [GPT-3.5] As duas meninas que foram tão amigas no passado hoje já não têm mais contato. [GPT-3.5]

Table B3. An example of an application of LLMs for GEC that goes beyond error correction capabilities. Using chain-of-thought, we ask the LLM to 1) check whether the phrase has grammatical errors, 2) correct the phrase in case of grammatical errors, and 3) explain why the phrase is wrong in case of grammatical errors.

PROMPT

`You are an agent capable of correcting phrases to grammatically correct Brazilian Portuguese.`
 	
`Think step by step in the following format:`
`1. Does the phrase have any grammatical errors?`
`2. In case of grammatical errors, correct the phrase.`
`3. Explain why the phrase is wrong.`

PHRASE: *"João sempre tem um caloroso comprimento para oferecer quando nos encontramos por acaso."*

`1. Yes, the phrase has a grammatical error.`
`2. The correct version would be: "João sempre tem um caloroso cumprimento para oferecer quando nos encontramos por acaso."`
`3. The error in the original phrase was the misspelling of "cumprimento" as "comprimento". "Cumprimento" means "greeting" or "salute", while "comprimento" means "length". Therefore, the original phrase didn't make sense in the context of greeting someone.`

PHRASE: *"A esposa e o marido tem muito dinheiro."*

`1. Yes, the phrase has a grammatical error.`
`2. The correct version would be: "A esposa e o marido têm muito dinheiro."`
`3. The verb "ter" (to have) is conjugated in the third person plural ("têm") to agree with the subjects "a esposa e o marido" (the wife and the husband), which are also in the third person plural. In the original sentence, the verb "tem" is used in the third person singular form, which does not agree with the subjects of the sentence.`