LLM-based post-editing as reference-free GEC evaluation

Robert Östling
Stockholm University
Sweden
robert@ling.su.se

Murathan Kurfalı
RISE Research Institutes of Sweden
Sweden
murathan.kurfali@ri.se

Andrew Caines
ALTA Institute & Computer Laboratory
University of Cambridge, U.K.
andrew.caines@cl.cam.ac.uk

Abstract

Evaluation of Grammatical Error Correction (GEC) systems is becoming increasingly challenging as the quality of such systems increases and traditional automatic metrics fail to adequately capture such nuances as fluency versus minimal edits, alternative valid corrections compared to the 'ground truth', and the difference between corrections that are useful in a language learning scenario versus those preferred by native readers. Previous work has suggested using human post-editing of GEC system outputs, but this is very labor-intensive. We investigate the use of Large Language Models (LLMs) as post-editors of English and Swedish texts, and perform a meta-analysis of a range of different evaluation setups using a set of recent GEC systems. We find that for the two languages studied in our work, automatic evaluation based on post-editing agrees well with both human post-editing and direct human rating of GEC systems. Furthermore, we find that a simple n-gram overlap metric is sufficient to measure post-editing distance, and that including human references when prompting the LLMs generally does not improve agreement with human ratings. The resulting evaluation metric is reference-free and requires no language-specific training or additional resources beyond an LLM capable of handling the given language.

1 Introduction

Grammatical Error Correction (GEC) is an important technology for supporting native and non-native writers, and supporting the development of language learners (for a recent survey see, for instance, Bryant et al., 2023). In recent years, neural networks and in particular Large Language Models (LLMs) have led to rapid improvements in the accuracy of such systems, but these developments have made apparent the difficulty of efficiently evaluating such systems.

For the most part, reference-based metrics have been used for the evaluation of GEC. These metrics depend upon human-created reference corrections and either rely on text similarity measures similar to those used in Machine Translation – examples include GLEU (Napoles et al., 2015) and GREEN (Koyama et al., 2024) – or on comparing the edits made by the GEC system with those by the human; for instance, M^2 (Dahlmeier and Ng, 2012) and ERRANT (Bryant et al., 2017). These reference-based metrics have been shown to correlate less well with human quality estimates than other approaches, in particular with recent neural GEC systems (Kobayashi et al., 2024). In addition, the manual process of creating references is time-consuming.

Reference-free metrics, typically based on neural models, have been proposed as an alternative, but these tend to either be complex and requiring additional (language-specific) training data (Yoshimura et al., 2020; Maeda et al., 2022), or to be simplistic but may correlate relatively poorly with human preferences (Islam and Magnani, 2021).

Östling et al. (2024) proposed using human post-editing to create one reference per GEC system output, and then use a text similarity metric between the system output and its post-edited version as a measure of GEC system quality. This was evaluated on a small number of GEC systems in Swedish, so it is unclear to what extent the resulting scores correlate with human preferences. In addition, human post-editing of every system output is a very time-consuming task. Our goal in this work is to investigate whether the human post-editing step can be performed by an LLM, and how the evaluation setup can be modified to achieve maximal correlation with human evaluation by either

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 213–224

post-editing, ranking, or direct scoring.

Our main research questions are:

- RQ1: does LLM-based post-editing provide a scoring of GEC systems that aligns with human preferences?
  (Answer: yes, there is a high level of agreement with different types of human quality assessments.)
- RQ2: how does the choice of text similarity metric affect post-editing based GEC evaluation?
  (Answer: Levenshtein distance as used in previous work is sub-optimal, chrF++ is good but overkill; use character bag-of-6-gram overlap instead.)
- RQ3: what difference does it make if human references are provided to the LLMs while performing post-editing?
  (Answer: in general the best method is to use only the original sentence + system output, but peculiarities in some datasets affect this outcome.)
- RQ4: how does LLM-based post-editing compare to human post-editing for GEC system evaluation?
  (Answer: they generally agree very well, but LLMs make somewhat more changes and have a considerably lower proportion of completely unchanged sentences.)

2 Related Work

Grammatical error correction has a long history as an area of research (Bryant et al., 2023). It has also featured in various shared tasks over the years (e.g., Ng et al., 2014; Bryant et al., 2019; Masciolini et al., 2025). Since statistical approaches to GEC were widely adopted, the best-performing systems involve supervised models trained on annotated corpora: usually involving sequence-to-sequence models (e.g., Rothe et al., 2021) or pipeline systems based on sequence tagging (e.g., Omelianchuk et al., 2020).

According to recent research, LLMs do not outperform these supervised GEC systems on every benchmark, at least for English (Loem et al., 2023; Davis et al., 2024). Instead, it has been shown that they can potentially improve the recall of GEC models in an ensemble setting (Omelianchuk et al., 2024). Moreover, given the increasing use of LLMs as judges, in this work we investigate to what extent LLMs can be used for GEC evaluation which, along with the availability of high quality annotated data, is a bottleneck to progress in GEC (Kobayashi et al., 2024).

Current metrics are either reference-based or reference-free, meaning that they do or do not, respectively, depend upon 'ground truth' corrections. The most widely-used reference-based metrics are precision, recall and $F_{0.5}$ – most often obtained from the M^2 scorer (Dahlmeier and Ng, 2012) or with ERRANT (Bryant et al., 2017) – along with GLEU, derived from the BLEU score commonly used in machine translation (Napoles et al., 2015). However, there is often more than one possible way to correct a grammatical error, and even with multiple annotations it is difficult to cover all possibilities in reference-based approaches.

Examples of reference-free metrics include the Scribendi Score (Islam and Magnani, 2021) and IMPARA (Maeda et al., 2022). The former may involve any LLM, in principle, whilst the latter was implemented using BERT (Devlin et al., 2019). However, the reliance on language models for reference-free metrics means that they tend to be biased towards fluency corrections over minimal edits which stay closer to the original text formulation but may not be recognized as improvements by the language models. Fluent corrections are usually preferable from a readability and naturalness perspective, but it is arguable from a pedagogical standpoint that it is better to in fact offer minimal edits as feedback to human learners rather than error avoidance strategies (Sakaguchi et al., 2016; Caines et al., 2023; Mita et al., 2024).

Nevertheless, the reliance on ground truth references remains a limiting factor in evaluation of GEC systems on new data. If it can be shown that LLMs can be reliably put to use as GEC post-editors, for the purpose of evaluation, correlating well with human judgements, it would release the pressure on the GEC bottleneck somewhat. Östling et al. (2024) examine the feasibility of post-editing based evaluation with Swedish GEC data and perform direct scoring as well as post-editing of the outputs of three different GEC systems and two fluency-edited references. They find that post-editing distance correlates strongly with the scores assigned by the annotator, but the small sample of GEC systems limits the range of conclusions that they are able to draw. Additionally, their annotation procedure is fully manual and would be difficult to

scale up.

3 Data

Kobayashi et al. (2024) performed a meta-evaluation of 12 recent English GEC systems, and published the SEEDA dataset of GEC system outputs and human rankings of sentences from these outputs. We use this dataset because it contains a sufficient number of GEC systems to compute reasonably reliable correlations between a given GEC evaluation metric and the human assessments. In addition to system outputs of 12 modern GEC systems, it also includes the original uncorrected sentences (INPUT) and two human-created reference, one with minimal edits (REF-M) and one edited for fluency (REF-F).

Östling et al. (2024) published human annotations with post-edited versions of 3 Swedish GEC systems as well as the original uncorrected sentences (INPUT) and three human-created references, one with minimal edits (REF-M) and two edited for fluency. The GEC system outputs and the fluency-edited references are annotated with scores for grammaticality, fluency and meaning preservation, and post-edits to achieve perfect scores in these three assessment dimensions. We include the Swedish data for two main purposes: to allow direct comparisons between human and LLM post-edits, and to verify that the proposed method can be applied to languages other than English given a suitable LLM.

4 Method

We have several different recent LLMs perform post-editing of GEC system outputs from the datasets of Kobayashi et al. (2024) in English, and Östling et al. (2024) in Swedish.[1] We use Gemma 2 in several sizes (2 billion parameters, 9B, 27B) (Gemma Team et al., 2024), Gemma 3 27B (Gemma Team et al., 2025), Llama 3.1 8B (Grattafiori et al., 2024), Mistral Small 24B (Jiang et al., 2023), Qwen 2.5 32B (Bai et al., 2023), and Command A-111B (Cohere, 2025).

For each LLM, we try each combination of the following two parameters:

- Semantic grounding. In order to ensure that the post-editing does not diverge from the semantics of the original text, we include four (English) or three (Swedish) types of semantic grounding. In all cases the GEC system output is provided in the prompt.
  - None. Only the system output is provided in the prompt.
  - INPUT. The GEC system input (original text) is included.
  - REF-M. A human minimal edits reference is included.
  - REF-F. A human fluency edited reference is included (English data only).
- Similarity metric. Following Östling et al. (2024) we use Normalized Levenshtein distance as one metric, and add two n-gram-similarity-based metrics.
  - Normalized Levenshtein Similarity, which is identical to Normalized Levenshtein Distance apart from the direction (higher is better):

$$S(a, b) = 1 - L(a, b)/\max(|a|, |b|)$$

  - chrF++ (Popović, 2017), which in our setting computes the mean F_2 score over word bigram and character 6-gram precision and recall. Unlike the other metrics, this is asymmetric and we treat the post-edited text as the reference.
  - Character 6-gram bag-of-n-grams overlap, a symmetric measure of similarity:

$$S(a, b) = |N_a^6 \cap N_b^6|/|N_a^6 \cup N_b^6|$$

    where N_s^6 is the set of character 6-grams (including spaces) for string s.

Because it is difficult to justify a full parameter search using the very largest model (GPT-4o[2]), we obtain post-edits only for the setting where the original sentence is used as semantic grounding (INPUT), since this was the most promising configuration in preliminary experiments. For the Swedish part we also restrict the set of LLMs used to some of the models that obtained the most promising results on the English data, due to time and data licensing constraints.[3]

[1] Prompts are given in Appendix A.

[2] The actual number of parameters has not been published for it or its smaller version GPT-4o-mini, but we see a limited value in exploring the full set of parameters for these models.

[3] The current license of the Swedish data does not permit the use of OpenAI API.

Post-editor	r	ρ
Gemma 2-2B	0.69	0.77
Gemma 2-9B	**0.95**	**0.92**
Gemma 2-27B	0.79	0.56
Gemma 3-27B	0.82	0.67
Llama 3.1-8B	0.90	0.83
Mistral Small 24B	**0.95**	0.91
Qwen 2.5-32B	0.81	0.68
Command A-111B	**0.95**	0.89

Table 1: System-level correlations between post-edit distance and human ratings, averaged over all similarity metrics, semantic grounding options, and human ratings. Here and below boldface is used as a visual aid to identify the highest values.

Similarity metric	r	ρ
Levenshtein	0.74	0.63
6-gram overlap	**0.92**	**0.85**
chrF++	0.91	**0.85**

Table 2: Mean system-level correlations between post-edit distance and human ratings, averaged over all LLM post-editors, semantic grounding options, and human ratings.

Semantic grounding	r	ρ
None	0.83	0.66
INPUT	**0.88**	**0.87**
REF-M	0.83	0.76
REF-F	**0.88**	0.82

Table 3: Mean system-level correlations between post-edit distance and human ratings, averaged over all LLM post-editors, similarity metrics, and human ratings.

Human rating	r	ρ
EW/edit	0.84	0.77
EW/sentence	0.85	**0.79**
TS/edit	**0.87**	0.77
TS/sentence	**0.87**	**0.79**

Table 4: Mean system-level correlations between post-edit distance and human ratings, averaged over all LLM post-editors, similarity metrics, and semantic grounding options. The four human rating references are computed using Expected Wins (EW) or TrueSkill (TS) from sentence-level rankings that are either edit-based or sentence-based.

For English, we follow Kobayashi et al. (2024) and compute correlations (Pearson r and Spearman ρ) to human annotations on the system level.[4] These are derived in two different types of annotation (edit-based or sentence-based comparisons), using two different methods (TrueSkill and Expected Wins) of summarizing the rankings into numeric scores, resulting in four different system-level references. To avoid making arbitrary decisions on which of these to prefer, and to increase the reliability of the results, we consistently use means over all of these four except in Table 4 where we investigate the effect of the human system-level score type and find that it is relatively small. For the sentence level evaluations we use Kendall τ, as computed by the software published by Kobayashi et al. (2024), for comparing to human sentence-level rankings.

For the Swedish data the available annotations are different, compared to English. Instead of rankings of system outputs, each system output has been annotated for grammaticality, fluency and meaning preservation. If any of these are annotated with less than a perfect score (4 on a scale 1–4), there is also a post-edited version of the system output with the goal of performing minimal editing to achieve full scores on all three properties. Since there are only three GEC system outputs and two human references included in the data, we do not consider it meaningful to perform a system-level evaluation as in the English data. Instead, we use Spearman's ρ to compare the post-edit score between the human annotator and each LLM. We also compare the LLM post-edit scores to the mean of the human annotator's grammaticality, fluency and meaning preservation scores, which we use as a general measure of the quality of that particular correction.

5 Results and Discussion

5.1 Overall agreement with human rankings

In order to see whether LLM-based post-editing provides a scoring of GEC systems that aligns with human preferences (RQ1), we begin by applying the meta-evaluation framework of Kobayashi et al. (2024). Because our proposed evaluation setup has several hyperparameters and only 15 system outputs[5] to measure correlations with, we search

[4] Sentence 22 of the REF-M file in the SEEDA dataset is empty. We handle this by arbitrarily giving this sentence a score of 0 for all similarity metrics.

[5] Whenever the semantic grounding uses one of the human references (REF-M or REF-F), that reference is excluded from computing the correlation and only the remaining 14 system outputs are used. Note that unless stated otherwise, we use the term "system output" to also include the human-created

	Spearman ρ					Pearson r				
LLM	Base	None	INPUT	REF-M	REF-F	Base	None	INPUT	REF-M	REF-F
Gemma 2-2B	-0.28	0.94	0.61	0.74	0.93	-0.64	0.97	0.58	0.72	0.96
Gemma 2-9B	0.35	0.91	0.95	0.94	0.94	-0.18	0.96	0.97	0.97	0.96
Gemma 2-27B	0.55	0.68	0.92	0.83	0.60	0.05	0.87	0.97	0.90	0.83
Gemma 3-27B	0.46	0.42	0.94	0.83	0.89	-0.15	0.76	0.97	0.91	0.93
Llama 3.1-8B	0.14	0.95	0.93	0.92	0.92	-0.41	0.97	0.95	**0.98**	**0.98**
Mistral Small 24B	0.29	0.91	**0.96**	0.94	0.95	-0.27	0.97	**0.98**	**0.98**	0.95
Qwen 2.5-32B	0.56	0.44	0.94	0.89	0.83	-0.00	0.75	0.96	0.89	0.92
Command A-111B	0.48	0.87	0.95	0.93	0.93	-0.06	0.95	**0.98**	**0.98**	0.94
GPT-4o	–	–	**0.96**	–	–	–	–	**0.98**	–	–
GPT-4o-mini	–	–	**0.96**	–	–	–	–	0.97	–	–

Table 5: Mean system-level correlations between post-edit distance and human ratings, per LLM and semantic grounding option, always using 6-gram overlap and averaging over human ratings.

	Sentence-based					Edit-based				
LLM	Base	None	INPUT	REF-M	REF-F	Base	None	INPUT	REF-M	REF-F
Gemma 2-2B	-0.21	0.32	0.18	0.23	0.32	-0.13	0.35	0.21	0.27	0.33
Gemma 2-9B	0.10	0.36	**0.54**	0.41	0.38	0.15	0.35	0.52	0.41	0.39
Gemma 2-27B	0.21	0.18	0.42	0.25	0.15	0.26	0.18	0.41	0.25	0.20
Gemma 3-27B	0.11	0.14	0.47	0.28	0.29	0.20	0.11	0.45	0.27	0.24
Llama 3.1-8B	-0.01	0.33	0.36	0.30	0.31	0.06	0.35	0.39	0.34	0.34
Mistral Small 24B	0.08	0.35	0.48	0.38	0.39	0.18	0.33	0.50	0.38	0.33
Qwen 2.5-32B	0.21	0.14	0.45	0.23	0.26	0.24	0.16	0.45	0.22	0.23
Command A-111B	0.19	0.38	0.46	0.37	0.38	0.22	0.37	0.47	0.37	0.39
GPT-4o	–	–	**0.54**	–	–	–	–	**0.55**	–	–
GPT-4o-mini	–	–	0.46	–	–	–	–	0.46	–	–

Table 6: Mean sentence-level Kendall τ between post-edit distance and human ratings, per LLM and semantic grounding option, always using 6-gram overlap.

through each parameter independently taking the averages over all other parameters in order to avoid overfitting. Averaged system-level correlations are presented in Table 1 (per LLM), Table 2 (per similarity metric), and Table 3 (per semantic grounding option). Additionally, we also present the averaged correlations per human rating setup (Table 4) and see that these are in general agreement with each other. In all other system-level evaluation results, we present averages over all four human rating setups to obtain more reliable estimates.

5.2 Effect of text similarity metric

Next, we turn to the question of how the text similarity metric used to compare the system output with its post-edited version affects the results (RQ2). Östling et al. (2024) used Normalized Levenshtein Distance with manual post-edits. We compute the its negated version (Normalized Levenshtein Similarity) along with two other options. The results are shown in Table 2, averaged over all other parameters. It is clear that Normalized Levenshtein Similarity is in fact sub-optimal, and that both of the other two metrics obtain correlations with human ratings that are considerably higher. In the following analysis we use 6-gram overlap, as it is simple and efficient to compute.

5.3 Effect of semantic grounding

To investigate whether the type of semantic grounding affects post-editing based evaluation (RQ3), we compute the correlations separately for the different types of semantic grounding (Table 5). There are pronounced differences between the various LLMs with respect to which type works best, but the overall trend is that adding human-written references typically does not improve the outcomes, and in most cases results in lower correlation with human ratings.

We have included a baseline (Base) consisting of a reference generated by the same LLM *without* access to the system output, using the LLM as a GEC system with access to the original text only.[6] This is done to exclude the possibility that the LLMs generate high-quality references and that post-editing is an unnecessary complication. However, the low correlation values for the baseline indicate that including the system output and performing post-editing is essential to the success of our method. Manual inspection indicates that REF-F and GPT-3.5, both of which contain a considerable amount of fluency edits, are generally rated poorly by the baseline.

LLM	S.G.	HP	HS
Gemma 2-9B	None	0.39	0.39
Gemma 2-9B	INPUT	0.28	0.26
Gemma 2-9B	REF-M	0.55	0.51
Mistral Small 24B	None	0.40	0.38
Mistral Small 24B	INPUT	0.30	0.30
Mistral Small 24B	REF-M	0.55	**0.51**
Qwen 2.5-32B	None	0.35	0.34
Qwen 2.5-32B	INPUT	0.34	0.34
Qwen 2.5-32B	REF-M	0.49	0.45
Command A-111B	None	0.49	0.46
Command A-111B	INPUT	0.40	0.39
Command A-111B	REF-M	**0.58**	0.53

Table 7: Spearman ρ between LLM post-edit score, and each of human post-edit (HP) and human score (HS, mean of grammaticality, fluency and meaning preservation scores). Scores from post-edits are defined as the 6-gram similarity to their respective system output. The correlation between HP and HS is 0.81. S.G. = semantic grounding.

It is also noteworthy that some of the highest system-level correlations are obtained by letting the smallest of the evaluated LLMs (Gemma 2-2B) post-edit the system output with only the system output and no semantic grounding, thus ignoring any possible semantic errors. In line with previous work (Yoshimura et al., 2020) which found that meaning preservation is not an important factor when trying to achieve high correlation to human ratings, this indicates that having even a modest-sized LLM perform conservative correction of the system output brings us to close agreement with human system-level ratings.

5.4 Sentence-level evaluation

We now turn from system-level to sentence-level evaluations. Following Kobayashi et al. (2024), we present the sentence-level agreement with human rating as Kendall τ values in Table 6. At this finer level of granularity, the differences between different metric parameters become apparent. Adding the original sentence as semantic grounding consistently improves the correlation with human assessments, while adding a human reference (REF-M or REF-F) shows no such tendency. Again, the baseline consistently has very low correlations.

references.

[6] Prompts are given in Appendix A.

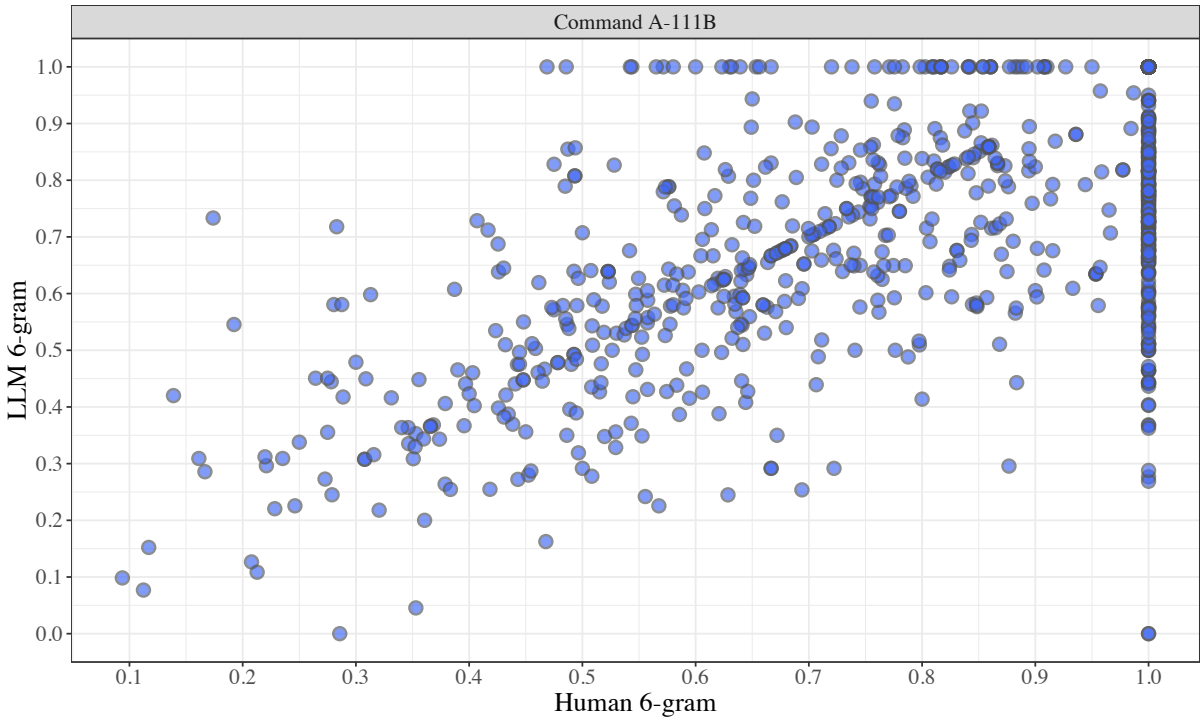

Figure 1: Scatter plot of character 6-gram overlap scores derived from human (x-axis) and LLM (y-axis) post-edits, in both cases using REF-M for semantic grounding. A score of 1 indicates that no changes were made during post-editing. The points are one-third transparent to avoid over-plotting.

5.5 Human vs. LLM post-edits

In order to investigate the relationship between post-edits made by humans and LLMs (RQ4), we use the Swedish data from Östling et al. (2024), where three GEC system outputs and two human references have been post-edited as well as rated for grammaticality, fluency and meaning preservation. We used a subset of the most promising LLMs to replicate the post-editing and allow direct comparisons between LLM and human post-edits. Table 7 presents correlations between LLM post-edit scores (using character 6-gram overlap) and human post-edit scores (also using character 6-gram overlap) as well as to the mean of the grammaticality, fluency and meaning preservation scores. The latter is used to approximate a direct assessment by the human annotator of the GEC system's output of that particular sentence.

In the human post-editing of Östling et al. (2024), a minimal edits reference (REF-M) was used for semantic grounding. As expected, we find that using this reference in the LLM prompt leads to higher correlation to both the human post-edit distance and the human annotated scores. Unlike for the English SEEDA data, using only the original sentence for semantic grounding (INPUT) leads to considerably lower correlations. We believe this to be due to the fact that the Swedish data consists of individual sentences in random order, and that only the creator of the REF-M reference has access to a wider context, while both the human and LLM post-editors lack any such context.

Figure 1 shows the 6-gram overlap scores assigned to each sentence from both the human post-editing and LLM post-editing. The LLM used was the one with the highest correlation to human post-editing scores (Cohere Command A-111B). We see that there is generally high agreement, as the $\rho = 0.58$ correlation indicates, but that there are some clear differences. The human post-editor frequently (46%) leaves the sentence unchanged, whereas the LLM does this less often (27%). The same tendency of the human post-editor being more reluctant to change is reflected in the mean overlap scores: 0.81 (SD 0.22) for the human, compared to 0.74 (SD 0.22) for the LLM, meaning that on the whole the human annotators post-edited less of the system output than the LLMs did. A significant part of this difference is due to the cases where humans leave sentences unchanged, which is demonstrated by considering only sentences where both the human and the LLM actually perform

some edits. In this case, the correlation between the 6-gram overlap scores increases to $\rho = 0.67$ for the same model.

5.6 LLMs as GEC systems and post-editors

An important question[7] is whether LLMs can be expected to post-edit the output of LLM-based systems, and if it would not be better to simply use the LLMs as GEC systems to begin with.

Our method is based on the assumption that an LLM is capable enough to post-edit the output of even the best GEC systems under evaluation. We have found this to be the case in our evaluation where even the best LLM-based systems undergo significant post-editing during evaluation. Furthermore, we argue that the availability of an LLM with sufficiently high capability is a realistic assumption in a practical setting, since considerably more computation can be spent on GEC evaluation (which will be run once or a few times) than on actual deployed GEC systems.

It is also important to note that GEC evaluations will also be needed for non-LLM based systems. Kobayashi et al. (2024) worked with 12 systems to carry out English GEC for the SEEDA dataset. Östling et al. (2024) worked with 3 systems for Swedish GEC of essays in the SweLL dataset (Volodina et al., 2019). In both cases the systems include both supervised and unsupervised approaches, for instance involving machine translation, sequence tagging and few-shot prompting of LLMs. That is, we do evaluate both non-LLM and LLM systems for GEC in this work.

6 Conclusions

We find that LLMs can be used as very effective evaluation tools for GEC systems, by asking them to post-edit system outputs and using a simple string similarity metric (character 6-gram overlap) to measure the amount of editing needed to go from the GEC system's output to a version considered by the LLM to be fully grammatical and fluent, while completely preserving the meaning expressed in the original. Even relatively small LLMs (such as Gemma 2-2B) can perform this task well enough to achieve nearly perfect correlation with human ratings at the system level. However, the picture is different when the GEC system output is assessed on the level of individual sentences, with considerable variation between LLMs in the ability to predict the human assessment of that sentence.

While we use the most recent publicly availabe GEC meta-evaluation dataset (Kobayashi et al., 2024), LLM-based GEC systems improve rapidly and an important question is to what extent LLM-based post editing is able to evaluate the output of the most capable LLMs. Answering this would require additional annotations that go beyond the scope of this work.

To summarize, we see several advantages of evaluation based on post-editing GEC system outputs by LLMs:

- High correlations with human direct assessment of GEC system quality, both at the system level and sentence (or document) level.
- Analyzing the post-edits provides an interpretable indication of the weaknesses of a particular GEC system, and this can be partly automated by tools such as ERRANT (Bryant et al., 2017). This contrasts with ranking-based evaluations like that recently proposed by Goto et al. (2025).
- Given a multilingual LLM, post-editing can handle multiple languages without requiring any additional language-specific resources or training.
- Unlike metrics that depend on having a large number of data points to average over (e.g., Islam and Magnani, 2021; Goto et al., 2025), post-editing distance can be estimated even on a single document without a sentence-aligned system output. It is thus suitable for document-level evaluations, as in Masciolini et al. (2025).

Fully exploring document-level multilingual evaluation would be an interesting direction of future work (Piotrowska, 2025). Note that in this work we have only worked at the sentence level, as has been conventional in GEC for the most part. However, in recent years there has been growing interest in document-level GEC, as well as evidence that the additional context can aid system performance on certain error types which relate to linguistic features above the sentence level (Yuan and Bryant, 2021; Mita et al., 2024; Masciolini et al., 2025).

[7] Raised by one of the anonymous reviewers.

Limitations

This paper on GEC is limited in the sense that we work with only 2 languages (English and Swedish) and findings for other languages may vary from those reported here. Annotated data for GEC are costly to build and therefore hard to come by: the datasets we work with in this paper are relatively small, compared to some corpora used in other areas of NLP. In addition the correction of grammatical errors is to some extent subjective, and an estimation without full access to the authors' original intentions. However, this limitation is a factor for all working on GEC.

LLMs have proven to be highly effective for a number of NLP tasks. In this paper we show that they are not necessarily state-of-the-art at the GEC task itself, but may be sufficiently accurate on the GEC post-editing task. This finding is limited by the continued availability of high quality open-weights LLMs, sufficient computing resources for those conducting research to be able to use the LLMs for inference, and the fact that we have only evaluated their performance on two languages. However, in principle, many LLMs have highly multilingual capabilities, and we expect that the outcomes reported here will hold for many other languages.

Acknowledgments

We thank the three anonymous reviewers for their valuable comments.

This work is partly funded by the Swedish national research infrastructure Språkbanken, jointly financially supported by the Swedish Research Council (2018–2028; grants 2017-00626 and 2023-00161) and the 10 participating partner institutions. The second author is partially supported by the Swedish Research Council under grant agreement no. 2024-01506. The third author is supported by Cambridge University Press & Assessment.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, 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*, pages 643–701.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology. In *Proceedings of the Empowering Education with LLMs – the Next-Gen Interface and Content Generation Workshop at AIED*.

Cohere. 2025. Command A: An enterprise-ready large language model.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, 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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey

Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025. Rethinking evaluation metrics for grammatical error correction: Why use a different evaluation process than human? *Preprint*, arXiv:2502.09416.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

Shota Koyama, Ryo Nagata, Hiroya Takamura, and Naoaki Okazaki. 2024. n-gram F-score for evaluating grammatical error correction. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 303–313, Tokyo, Japan. Association for Computational Linguistics.

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.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia. University of Tartu Library.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2024. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 251–265, Mexico City, Mexico. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. 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 Skurzhanskyi. 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.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. 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.

Emilia Piotrowska. 2025. Multilingual document-level gec evaluation. Bachelor's thesis, Department of Linguistics, Stockholm University.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and 1 others. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Online. Association for Computational Linguistics.

A Prompts

In this appendix, we present the prompts used across different experiments. We use a total of three prompt templates: one for the baseline results where LLMs are applied to GEC tasks, and two for the post-editing experiments—one without a semantic grounding sentence and one with. The same prompt structure is used for all three types of semantic grounding.

Prompt for GEC baseline

> Reply with a corrected version of the input sentence with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.
>
> Instructions:
>
> 1. Return ONLY the corrected sentence.
> 2. Wrap the corrected sentence in `<corrected>` and `</corrected>` tags.
> 3. Do NOT include any explanations, extra text, or formatting.
>
> Example:
>
> > <corrected>This is your corrected sentence.</corrected>
>
> Input sentence: {sentence}
> Output:

Prompt for post-editing without semantic grounding

> Please make minimal modifications to the given sentence to achieve all of the properties below:
>
> - Perfect grammaticality: The sentence is native-sounding. It has no grammatical errors, but may contain very minor typographical and/or collocation errors.
> - Perfect fluency: The sentence sounds extremely natural and native-like.
> - Same language: The sentence must remain in the same language as the original (do not translate or change language).
>
> Instructions:
>
> 1. Return ONLY the corrected sentence.
> 2. Wrap the corrected sentence in `<corrected>` and `</corrected>` tags.
> 3. If the original sentence is already perfect, return it AS IS inside the `<corrected>` tags.
> 4. Do NOT include any explanations, extra text, or formatting.
>
> Example output format:
>
> > <corrected>Your corrected sentence here.</corrected>
>
> Sentence: {sentence}
> Output:

Prompt for post-editing with semantic grounding

> Please make minimal modifications to the given sentence to achieve all of the properties below:
>
> - Perfect grammaticality: The sentence is native-sounding. It has no grammatical errors, but may contain very minor typographical and/or collocation errors.
> - Perfect fluency: The sentence sounds extremely natural and native-like.
>
> Instructions:
>
> 1. Return ONLY the corrected sentence.

2. Wrap the corrected sentence in `<corrected>` and `</corrected>` tags.
3. Ensure that the corrected sentence preserves the meaning of the reference sentence provided below. The reference may contain grammatical errors — it is for semantic grounding only.
4. If the original sentence is already perfect, return it AS IS inside the `<corrected>` tags.
5. Do NOT include any explanations, extra text, or formatting.

Example output format:

> <corrected>Your corrected sentence here.</corrected>

Sentence: {sentence}
Reference (for meaning preservation only): {reference sentence}
Output: