

AMU-GEC: A Polish Native Speaker Grammatical Error Correction Benchmark

Ryszard Staruch^{1,2}

ryszard.staruch@amu.edu.pl

Jarosław Liberek^{1,2}

jaroslaw.liberek@amu.edu.pl

Krzysztof Jassem^{1,2}

krzysztof.jassem@amu.edu.pl

Filip Graliński¹

filip.gralinski@amu.edu.pl

Michalina Klimczak^{1,2}

mickli3@st.amu.edu.pl

Sonia Ołubczyńska^{1,2}

sonolu@st.amu.edu.pl

¹Adam Mickiewicz University ²Center for Artificial Intelligence

Abstract

For many languages, grammatical error correction (GEC) remains underexplored due to the lack of dedicated datasets and evaluation benchmarks. Moreover, many existing GEC datasets are based on texts written by second-language learners, which narrows the research focus to specific error types. To address these gaps, we introduce a new Polish GEC benchmark, AMU-GEC, built from texts written by native Polish speakers and manually annotated by expert linguists. The benchmark is released as a machine learning challenge ¹, without revealing the expected labels, to prevent contamination. To set baselines, we evaluate recent multilingual and Polish LLMs, finding that smaller models may perform better when trained on monolingual (Polish) data. We also conduct a human analysis of the best performing model to gain more insights about the model behavior when it is different from human annotations. The analysis shows that most of the changes suggested by the model do not introduce errors into the texts; instead, they typically either slightly improve text fluency or represent valid alternative corrections.

1 Introduction

Grammatical error correction (GEC) is the natural language processing task of identifying and correcting errors in texts. Over the years, the task has evolved to concern various types of linguistic errors, such as those that are strictly grammatical (e.g. subject-verb agreement), orthographic (e.g. spelling mistakes), semantic (e.g. wrong word

choice), syntactic (e.g. wrong preposition usage), or phraseological (e.g. inter-language calques).

In recent years, the field has seen substantial progress due to the effectiveness of neural network-based solutions, especially large language models (LLMs). These models have significantly improved correction quality for English (Coyne et al., 2023; Liang et al., 2025), enabling high-performing systems that support automated error correction in practical writing scenarios.

However, progress has been uneven across languages (Masciolini et al., 2025b). A recent initiative, the MultiGEC-2025 Shared Task (Masciolini et al., 2025a), aims to address this gap by organizing a multilingual grammatical error correction competition covering 12 European languages. Polish was not included in this Shared Task, and grammatical error correction for Polish remains an underexplored area.

Furthermore, most GEC research focuses on texts written by second-language learners. As a result, both the error types being corrected and the evaluation practices are tailored to the needs of this user group. Even if a system achieves promising performance on a benchmark based on L2 learner texts, it does not mean that it will perform equally well on texts written by more proficient speakers or in specialized domains. Qorib and Ng (2022) show the weaknesses of models that, according to the $F_{0.5}$ metric (which is one of the main GEC metrics), are supposed to outperform human annotations, and suggest that new test sets are needed for robust GEC evaluation.

To address these gaps, we introduce a new benchmark, AMU-GEC, based on Polish texts written by

¹<https://github.com/richardxoldman/AMU-GEC>

Dataset	Lang	Split	# examples	# sentences	# tokens	# references
JFLEG	en	dev	754	754	14.0k	4
JFLEG	en	test	747	747	14.1k	4
CoNLL-2014	en	test	1.3k	1.3k	30.1k	2
W&I-LOCNESS	en	dev	4.4k	4.4k	87.0k	1
W&I-LOCNESS	en	test	4.5k	4.5k	85.7k	5
CCTC-H	zh	test	200	3.8k	157k	1
QALB-14	ar	dev	1k	1k	53.8k	1
QALB-14	ar	test	948	948	51.3k	1
AMU-GEC (ours)	pl	dev	304	913	15.0k	2
AMU-GEC (ours)	pl	test	500	1545	24.6k	2

Table 1: Statistics on different GEC benchmarks. # References indicates the number of reference corrections provided per input sentence.

native speakers and corrected by expert linguists. We believe that such a benchmark is a crucial resource that opens many paths for Polish GEC research. This work is a collaboration between computer scientists and Polish linguists, aiming to provide AI researchers with a benchmark that includes up-to-date, professionally annotated real-world examples.

In summary, our main contributions are as follows:

- The first Polish GEC benchmark based on texts written by native speakers and corrected by experts.
- An extensive evaluation of recent LLMs and the rule-based LanguageTool system on the new benchmark.
- A human analysis of the best-performing model to gain a better understanding of its performance.

2 Related Work

2.1 Grammatical Error Correction Benchmarks

English GEC benchmarks such as the CoNLL-2014 Shared Task (Ng et al., 2014) and the BEA-2019 Shared Task (Bryant et al., 2019) have been widely used for error correction system evaluation over the past decade. For fluency correction systems, the most common benchmark is JFLEG (Napoles et al., 2017). These benchmarks are based on texts written by second-language learners and reflect the types of grammatical errors typical for that user group.

There are also datasets from other domains. The GMEG dataset (Napoles et al., 2019) contains texts from formal native writing on Wikipedia and informal native writing from Yahoo Answers. The CWEB dataset (Flachs et al., 2020) was created from randomly sampled paragraphs from Common Crawl and corrected by expert annotators. Although these corpora go beyond the language-learner domain, they have attracted little interest and do not serve as standard English GEC benchmarks.

There are also benchmarks for other languages (Mohit et al., 2014; Zhang et al., 2023; Yoon et al., 2023; Rozovskaya and Roth, 2019; Syvokon et al., 2023), which expand the availability of resources for multilingual GEC. Table 1 shows statistics on various existing GEC benchmarks. While these datasets differ in terms of the language proficiency of their authors, the works introducing them rarely discuss the rationale behind their choice of evaluation metric, instead defaulting to the standard metrics used in the field. As a result, most of these benchmarks rely on the $F_{0.5}$ or $F_{1.0}$ score as the primary metric.

An exception is the work by Wang et al. (2022), who introduce a Chinese GEC dataset based on texts written by native speakers. They choose $F_{2.0}$ as the main evaluation metric, arguing that a GEC system designed for native speakers should prioritize recall over precision.

While recent progress has expanded GEC research to new languages and domains, many languages and application areas remain underexplored, leaving substantial room for future work.

Source text: Kiedy urodziłam wszystko się zaczęło. Serce wali mi jak oszalone ciśnienie skaka i chodzę jak pijana.
Target text: Kiedy urodziłam, wszystko się zaczęło. Serce wali mi jak oszalone, ciśnienie skacze i chodzę jak pijana.
English translation: When I gave birth, everything started. My heart was pounding like crazy, my blood pressure was jumping and I was walking around like a drunk.
Explanation: In the original text two commas are missing and the verb “skakać” (to jump) is colloquially and incorrectly inflected as “skaka” instead of “skacze”.
Source text: Zarabiać więcej – sposób na inflację, na którą decyduje się coraz więcej osób.
Target text: Zarabiać więcej – sposób na inflację, na który decyduje się coraz więcej osób.
English translation: Earn more – a way of dealing with inflation that more and more people are choosing.
Explanation: In the original text the relative pronoun “którą” (which) is a reference to the feminine-gender word “inflacja” (inflation) instead of the masculine-gender word “sposób” (a way). The correct pronoun is “który” (masculine form) to make the sentence semantically correct.
Source text: Były wspomnienia, wiersze dedykowane dla Nadziei Artymowicz oraz utwory poetki recytowane i wyśpiewywane przez artystów.
Target text: Były wspomnienia, wiersze dedykowane o Nadziei Artymowicz oraz utwory poetki recytowane i wyśpiewywane przez artystów.
English translation: There were reminiscences, poems dedicated to Nadzieja Artymowicz and the poet’s works recited and sung by artists.
Explanation: The original text contains an unnecessary preposition “dla” (for). In English, it is correct to use the preposition “to” after the word “dedicated”, but in Polish, the correct usage is to inflect the noun directly without adding the preposition “dla”.

Table 2: Examples from the development part of the AMU-GEC benchmark.

2.2 Evaluation Metrics

The F_β score remains the standard metric for the GEC task, with $\beta = 0.5$ being the most commonly used value. This adaptation, which places greater weight on precision, was introduced in the CoNLL-2014 Shared Task, based on texts written by second-language learners, with the argument that providing an erroneous correction is more harmful than not suggesting a change (Ng et al., 2014). Together with precision and recall, it provides a clear indication of how well system predictions match the reference annotations. These metrics are calculated using tools such as the M² Scorer (Dahlmeier and Ng, 2012) or the Error Annotation Toolkit (ER-RANT) (Bryant et al., 2017), which automatically extract edits between original and corrected texts.

Several alternative metrics have been proposed, including GLEU (Napoles et al., 2015), Scribendi (Islam and Magnani, 2021), PT-M² (Gong et al., 2022), IMPARA (Maeda et al., 2022), and its GED-oriented adaptation, IMPARA-GED (Sakai et al., 2025). However, their correlation with human judgment has typically been evaluated on minimal-edit GEC datasets, using system outputs that are sub-

stantially less effective than current state-of-the-art models (Grundkiewicz et al., 2015). This may have led to an outdated or biased perception of the reliability of these metrics. Notably, even some recent work introducing a meta-evaluation benchmark for English GEC continues to use the same dataset (Kobayashi et al., 2024). These limitations make it difficult to establish a universally accepted evaluation metric.

To support consistent evaluation practices across different metrics, a new tool called GEC-metrics (Goto et al., 2025) was recently introduced. The tool enables GEC systems to be evaluated using the most common metrics while ensuring that the evaluation process is reproducible.

2.3 Polish GEC

To our knowledge, only two GEC datasets are available for Polish: the Ermlab dataset² and the PIEWi dataset (Grundkiewicz, 2013). However, our analysis indicates that the overall quality of these resources is limited. The Ermlab dataset contains

²<https://github.com/Ermlab/polish-gec-dataset>

Set	P	R	$F_{0.5}$	$F_{1.0}$
Dev IAA	84.69	86.55	85.06	85.61
Test IAA	93.42	94.39	93.61	93.90

Table 3: Inter-annotation agreement between the two annotations of AMU-GEC benchmark. P stands for precision and R stands for recall.

errors introduced in a non-natural way (including some in the human-annotated portion), while the PIEWi dataset was automatically constructed from Wikipedia revision histories, making it a source of naturally occurring errors and corrections. However, the extraction process relied on rule-based filtering of Wikipedia edits, which introduces annotation noise and results in some incorrectly aligned or misannotated correction pairs. To date, no published research has used either dataset in experiments on Polish GEC. These datasets do not provide an explicit evaluation methodology, so they appear to be primarily suited for training rather than benchmarking.

A related initiative is the recently proposed benchmark for evaluating Polish linguistic and cultural competency (Dadas et al., 2026), which includes six categories, one of which is grammar. While the grammar tasks test LLM knowledge of Polish grammar, they do not directly evaluate grammatical error correction capabilities.

The most commonly used Polish GEC system is still the rule-based LanguageTool system (Naber, 2003; Miłkowski, 2010). The only publicly available neural model is a fine-tuned T5 model, which lacks any accompanying documentation³.

The lack of reliable datasets and systematic evaluation for Polish GEC highlights the need for a high-quality benchmark. Such a resource could support robust evaluation of GEC systems and facilitate future research in this area. This is particularly important given the morphological complexity of Polish, a highly inflected language with free word order and extensive case marking. These characteristics pose particular challenges for grammatical error correction, and our proposed benchmark is intended to address these issues by providing a reliable, linguistically grounded evaluation framework for Polish GEC.

³<https://huggingface.co/szymonj/polish-simple-error-correction>

Feature name	Dev	Test
Erroneous tokens	5.58%	5.61%
M operation rate	19.52%	18.10%
U operation rate	8.48%	7.25%
R operation rate	72.00%	74.65%
Punctuation error rate	38.20%	35.07%
Letter case error rate	1.85%	1.60%

Table 4: Detailed statistics for the AMU-GEC benchmark averaged across references. Operation labels are defined as: M – Missing, U – Unnecessary, R – Replacement. Punctuation and letter case error rates denote the percentage of all edits where the only difference between source and reference is punctuation or capitalization, respectively.

3 AMU-GEC Benchmark

3.1 Benchmark Origin and Linguistic Scope

Our benchmark is based on textual material collected by Polish linguists at Adam Mickiewicz University. As part of their work at the university, they run the Telephone Language Consulting Helpline, where people ask questions about the grammatical correctness of particular expressions. Usually, callers provide only a short phrase without a larger context, and so the linguists have to look for examples from real texts containing the error in question. The initial purpose of collecting these examples was to ensure that they clearly illustrated the reported grammatical error and helped to properly document the issue. This collection of examples was then used to create the AMU-GEC benchmark, whose origin makes it a unique and practical resource.

The material covers the most important types of linguistic errors: grammatical, including syntactic and inflectional errors, orthographic, phraseological, lexical, including cases influenced by incorrect phonetic realization, and punctuation errors. It includes both obvious mistakes and complex errors whose correction requires a deep understanding of the context. The knowledge gained by the linguists during their work at the Telephone Language Consulting Helpline provided them with a valuable native-speaker perspective on real-world language correctness issues.

The annotations in the AMU-GEC benchmark do not strictly adhere to prescriptive dictionary rules, as some of these sources have become outdated with language evolution. As a result, the examples have been carefully curated and selected

to reflect correct contemporary language use. Table 2 shows examples from the development part of the AMU-GEC benchmark.

3.2 Annotation Procedure and Inter-Annotation Agreement

Four Polish native speakers contributed to the original annotation process: two expert linguists and two non-expert native speakers. The examples were first corrected by the linguists, then reviewed from the perspective of non-expert language users to ensure that the annotations aligned not only with expert linguistic standards but also with the intuitions, expectations, and potential doubts of typical native speakers. Finally, the linguists performed a final verification step to produce the initial gold-standard annotations.

Subsequently, an additional expert linguist annotated the benchmark in a separate single-pass process. As a result, the benchmark contains two independent sets of annotations. We use these two annotation sets to estimate inter-annotation agreement, which reflects the degree of consistency between independently produced corrections.

We follow the approach of Bryant and Ng (2015) and compute precision, recall, and F-scores to measure inter-annotation agreement between the two annotations. Table 3 reports the agreement scores. The values are higher than those reported by Bryant and Ng (2015), which is expected given the proficiency level of the texts in our benchmarks: language learners tend to produce more errors, and these errors can often be corrected in multiple valid ways.

3.3 Benchmark Composition and Task Characteristics

The AMU-GEC benchmark consists of a development set containing 304 examples and a test set containing 500 examples. On average, each example contains 2.78 and 2.76 edit operations in the dev and test sets, respectively, where an edit operation corresponds to a correction such as replacing a word, inserting a missing word, or removing an unnecessary one. Each example is presented as a text pair consisting of the original text and its corrected version. Statistics for the benchmark are shown in Table 1 and Table 4.

We position AMU-GEC as an evaluation-only benchmark, as we believe that a strong evaluation resource is crucial for future research and can support multiple directions for developing training re-

sources and methods for Polish GEC. In particular, recent work has shown the effectiveness of LLMs for GEC in zero-shot and few-shot settings (Zeng et al., 2024), while other studies have demonstrated the potential of synthetic data generation for GEC (Grundkiewicz et al., 2019; Stahlberg and Kumar, 2024). We therefore view AMU-GEC as a reliable evaluation benchmark that can facilitate future work in these directions.

The AMU-GEC examples are mostly drawn from news articles, online forum comments, and blog posts. They cover a diverse range of topics, including health, politics, music, and sports. Since the texts were written by native speakers, they are generally fluent and contain far fewer errors than texts in language-learner GEC datasets (Flachs et al., 2020). For this reason, we frame the task as minimal-edit GEC.

However, this setting differs substantially from minimal-edit GEC for language learners. In learner texts, applying only minimal corrections may remove grammatical errors while leaving the resulting text somewhat unnatural or less than fully fluent, as discussed by Sakaguchi et al. (2016). In AMU-GEC, this issue is much less relevant: because the original texts are already fluent, correcting the remaining local errors should produce texts that are both error-free and fluent. Thus, the benchmark is intended to evaluate proofreading tools, whose goal is to correct errors rather than improve the overall fluency or style of the text.

Most GEC datasets are split at the sentence level. Due to the positive impact of larger context in GEC tasks (Yuan and Bryant, 2021), and the ability of modern solutions such as LLMs to handle longer inputs, we decided to divide our corpus into paragraphs, where a single paragraph may consist of one or more sentences. The average paragraph length is 295 characters for the development set and 300 characters for the test set. We consider this important because, in some cases, understanding the preceding and following parts of the text is essential for determining the appropriate correction for a given error. Although the number of examples is smaller than in JFLEG or CoNLL-2014, the total number of sentences and tokens is larger, which makes the size of AMU-GEC sufficient for reliable evaluation.

3.4 Data Release and Future Use

Due to the risk of data contamination in LLMs (Deng et al., 2024; Li et al., 2024), we do not re-

veal our labels, making our benchmark, both the development and test parts, available as a machine learning challenge on an ML evaluation platform. Because of this, only the text before the correction process is available. This ensures that the data with our annotations is protected from data leakage and that evaluation on any model remains fair.

4 System Evaluation on AMU-GEC

4.1 Methodology

To establish strong baselines for the benchmark, we evaluated recent open-source LLMs on the AMU-GEC benchmark using ERRANT via the GEC-metrics tool. We note, however, that although ERRANT supports span matching, its lack of Polish-specific error types limits the granularity of the analysis.

We used both multilingual and Polish models, including Bielik (Ociepa et al., 2025) and PLLuM (Kocoń et al., 2025). The evaluated models cover a wide range of sizes, but due to computational limitations, we tested only models with up to 70B parameters. The model selection criteria were based on general public interest, model size, country of origin and license openness. In addition, we evaluated the performance of LanguageTool, a rule-based solution that is still widely used by Polish users.

For each LLM, we used the same extensive prompt written in Polish, describing the minimal-edit GEC task, including the proper usage of Polish punctuation symbols. The prompt was provided in a zero-shot format, without any additional examples. The prompt text, and inference process is described in the Appendix.

4.2 Evaluation Metric Selection

Considering the domain and proficiency level represented in our benchmark, we use the F_1 -score rather than $F_{0.5}$ as the main evaluation metric. Since our benchmark goes beyond the traditional setting of assisting language learners, we do not give precision twice the weight of recall. In our setting, native speakers can often assess the correctness of changes suggested by a GEC system themselves, making balanced precision and recall more appropriate. We note that the use of the F_1 -score is well grounded in the GEC literature, as it was the main evaluation metric in several previous shared tasks (Ng et al., 2013; Mohit et al., 2014; Rozovskaya et al., 2015). We therefore consider this adaptation

important in light of the broader range of potential end users addressed by our benchmark.

4.3 Results

Table 5 presents the results on AMU-GEC. The LanguageTool system and the EuroLLM-1.7B-Instruct model achieve F_1 -scores below 30%, whereas larger models exceed 50%. This suggests that Polish GEC cannot be effectively addressed by relatively simple systems or smaller language models.

Among all evaluated models, gemma-4-31B-it achieves the highest F_1 -score of 74.12%, although several other models also obtain competitive scores above 60%. These results show that Polish GEC remains a challenging task for current LLMs, with considerable room for improvement, especially for smaller models.

The Bielik-11B-v2.6-Instruct and PLLuM-12b-nc-chat-250715 models, which were fine-tuned for Polish, achieve very promising scores, ranging from 65 to 68 percentage points. This suggests that adapting LLMs to a specific language can yield relatively strong performance given the model size.

It is also worth noting that some models tend to produce corrections with very high precision, resulting in $F_{0.5}$ scores above 75%. This highlights the distinct nature of the AMU-GEC benchmark, where systems need to strike an appropriate balance between precision and recall, in contrast to many existing benchmarks that rely on $F_{0.5}$ and, therefore, emphasize precision-oriented performance.

5 Analysis of the Best-Performing Model

When evaluating GEC models using an F-score metric, it is not possible to distinguish, for example, between changes that introduce errors and those that make only slight stylistic modifications and do not worsen the text. Therefore, to better understand the types of changes made by the best-performing model (gemma-4-31B-it) on our benchmark, we conducted a human analysis of its outputs. For all changes that do not match the human annotations, we define the following categories:

- **Missed correction:** the model failed to correct an error present in the source text.
- **Incorrect model change:** the model introduced an error into the text.
- **Valid alternative correction:** the model made a valid correction that differs from the

Model	Size	AMU-GEC dev			AMU-GEC test		
		P	R	F ₁	P	R	F ₁
Qwen3.5-4B	4B	57.28	54.81	56.02	54.40	53.21	53.80
Qwen3.5-9B	9B	57.62	61.42	59.46	59.23	59.59	59.41
Qwen3.5-27B	27B	58.78	70.51	64.11	59.46	71.13	64.77
gemma-4-E2B-it	5B	64.61	61.01	62.76	62.04	58.09	60.00
gemma-4-E4B-it	8B	67.78	66.42	67.09	65.54	61.62	63.52
gemma-4-31B-it	31B	69.67	78.47	73.81	69.21	79.77	74.12
EuroLLM-1.7B-Instruct	1.7B	40.58	16.64	23.61	45.80	15.46	23.11
EuroLLM-9B-Instruct	9B	81.07	44.77	57.69	72.65	38.84	50.62
EuroLLM-22B-Instruct-2512	22B	69.15	66.96	68.04	67.78	64.40	66.05
Bielik-PL-Minitron-7B-v3.0-Instruct	7B	45.76	55.79	50.28	44.58	50.72	47.45
Bielik-11B-v2.6-Instruct	11B	72.05	68.77	70.38	66.77	64.61	65.67
Bielik-11B-v3.0-Instruct	11B	66.92	65.08	65.99	59.25	59.07	59.16
Bielik-PL-11B-v3.0-Instruct	11B	45.31	61.93	52.34	39.89	59.82	47.86
PLLuM-12b-nc-instruct-250715	12B	90.38	48.58	63.19	86.95	43.64	58.11
PLLuM-12b-nc-chat-250715	12B	78.17	65.39	71.21	74.70	61.33	67.36
Llama-PLLuM-70B-instruct-250801	70B	90.39	63.10	74.32	84.89	56.94	68.16
Llama-PLLuM-70B-chat-250801	70B	84.47	67.27	74.89	80.36	61.34	69.57
LanguageTool (Polish rules)	–	39.35	14.23	20.90	32.84	11.77	17.33

Table 5: Results for selected models for the AMU-GEC benchmark. The part above the line shows results for multilingual models, and the part under the line shows results for models fine-tuned specifically for the Polish language.

Category	Count	Proportion
Missed Correction	181	37.63%
Incorrect Model Change	78	16.22%
Valid Alternative Correction	26	5.41%
Neutral / Out-of-scope Change	190	39.50%
Annotation Issue	6	1.25%
Total	481	100.0%

Table 6: Distribution of categories during human annotation for the gemma-4-31B-it model.

reference annotation.

- **Neutral / Out-of-scope Change:** the model made a neutral change that does not harm the text, but also does not correct an error, such as an editorial or stylistic change.
- **Annotation issue:** the model corrected an error that was not captured in the annotation, or it did not follow an annotation that appears to be incorrect.

The manual analysis was conducted by one of our expert linguists on the AMU-GEC test set. Table 6 presents the distribution of the above categories.

The most frequent category is **Neutral / Out-of-scope Change**, with 190 occurrences. These modifications can generally be viewed positively, as they tend to improve the overall fluency or stylistic quality of the text, even though they go beyond the strict scope of grammatical error correction. Such behavior is consistent with findings reported in previous studies (Coyne et al., 2023), which observed that large language models often introduce stylistic or paraphrastic edits in addition to direct error corrections.

The second most frequent category is **Missed Correction**, with 181 occurrences. This result indicates that the model still fails to identify a substantial number of existing errors. Although large language models demonstrate strong performance across many NLP tasks, these findings suggest that there is still considerable room for improvement in grammatical error correction systems for the Polish language.

The **Incorrect Model Change** category contains 78 occurrences. Importantly, this category accounts for only 27% of all model-generated changes that do not match the reference annotations. This observation suggests that the effective precision of the

model is higher (around 92%) than the standard precision metric would indicate, because conventional evaluation does not distinguish between genuinely harmful edits and benign modifications such as neutral changes or valid alternative corrections.

From the perspective of the annotator, the model especially struggles with the correct inflection of Polish names and surnames. In addition, many incorrect model changes are related to punctuation, indicating that these phenomena remain challenging even for large instruction-tuned models.

Overall, the analysis demonstrates that the model achieves very strong performance on the grammatical error correction task. Although the model still misses some errors and occasionally introduces incorrect changes, the majority of edits that diverge from the reference annotations do not negatively affect the grammatical correctness of the text.

6 Conclusions

Our work introduces a new Polish GEC benchmark based on texts written by native speakers. In contrast to existing benchmarks, our dataset consists of short paragraphs rather than single sentences. The corrected versions of the texts are not released to prevent contamination in potential solutions.

The evaluation process shows that current LLMs can produce promising results in a zero-shot setup, although there is still considerable room for improvement. One of the most notable findings is that models trained on monolingual Polish texts can perform well considering their size.

Our human analysis of the best performing model shows that most of the changes suggested by the model do not introduce errors to text; instead, these changes work as a slight fluency improvement which may positively impact the overall text. It shows that LLMs can work as promising GEC tools for the Polish language without fine-tuning for the task, but there is still room for improvement for both the model quality and its efficiency by using smaller models.

7 Future Work

As we establish a benchmark for Polish GEC with development and test data, future work could focus on creating training resources to support the development of more effective solutions beyond the provided baselines.

It would also be beneficial to evaluate whether lightweight AI models, after being fine-tuned for

the grammatical error correction task, can serve as effective proofreading tools for the Polish language without requiring high-performance GPUs or extensive VRAM.

Once such models are available, we plan to conduct an extensive human evaluation to gain deeper insights into their performance.

Limitations

Our dataset is relatively small and has two references for each paragraph, which may not cover all legitimate corrections that differ from the reference. Since we do not label error types, we lack insight into model performance on specific categories of errors, which limits detailed analysis.

We do not release training resources which would be beneficial to compare zero-shot LLMs performance with the fine-tuned models.

The dataset is restricted to written text from sources such as news articles, blogs, and forums, and may not fully represent other domains of the Polish language, such as academic or spoken language.

It would be beneficial to include the human evaluation of different LLMs to compare their performance and gain a deeper understanding of the behavior in the task.

References

- 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 and Hwee Tou Ng. 2015. [How far are we from fully automatic high quality grammatical error correction?](#) In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.
- 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. *Preprint*, arXiv:2303.14342.
- Śławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, and Rafał Poświata. 2026. [Evaluating polish linguistic and cultural competency in large language models](#). In *Artificial Intelligence and Soft Computing*, pages 60–71, Cham. Springer Nature Switzerland.
- 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.
- Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2024. [Investigating data contamination in modern benchmarks for large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.
- Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. [Grammatical error correction in low error density domains: A new benchmark and analyses](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8467–8478, Online. Association for Computational Linguistics.
- Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. [Revisiting grammatical error correction evaluation and beyond](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025. [gec-metrics: A unified library for grammatical error correction evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–534, Vienna, Austria. Association for Computational Linguistics.
- Roman Grundkiewicz. 2013. [Automatic extraction of Polish language errors from text edition history](#). In *Text, Speech, and Dialogue*, pages 129–136, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. [Human evaluation of grammatical error correction systems](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.
- 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.
- 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.
- Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyc, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, Bartłomiej Alberski, Tomasz Walkowiak, Aleksander Szczesny, Maciej Markiewicz, Tomasz Bernaś, Hubert Mazur, Kamil Żyta, Mateusz Tykierko, Grzegorz Chodak, Tomasz Kajdanowicz, Przemysław Kazienko, Agnieszka Karlińska, Karolina Seweryn, Anna Kołos, Maciej Chrabąszcz, Katarzyna Lorenc, Aleksandra Krasnodebska, Artur Wilczek, Katarzyna Dziewulska, Paula Betscher, Zofia Cieślińska, Katarzyna Kowol, Daria Mikoś, Maciej Trzcinski, Dawid Krutul, Marek Kozłowski, Śławomir Dadas, Rafał Poświata, Michał Perełkiewicz, Małgorzata Grębowiec, Maciej Kazuła, Marcin Białas, Roman Roszko, Danuta Roszko, Jurgita Vaičenonienė, Andrius Utkā, Paweł Levchuk, Paweł Kowalski, Irena Prawdzic-Jankowska, Maciej Ogrodniczuk, Monika Borys, Anna Bulińska, Wiktor Gumienna, Witold Kieraś, Dorota Komosińska, Katarzyna Krasnowska-Kieraś, Łukasz Kobylński, Martyna Lewandowska, Marek Łaziński, Mikołaj Łątkowski, Dawid Mastalerz, Beata Milewicz, Agnieszka Anna Mykowiecka, Angelika Peljak-Łapińska, Sandra Penno, Zuzanna Przybyś, Michał Rudolf, Piotr Rybak, Karolina Saputa, Aleksandra Tomaszewska, Aleksander Wawer, Marcin Woliński, Joanna Wołoszyn, Alina Wróblewska, Bartosz Żuk, Filip Żarnecki, Konrad Kaczyński, Anna Cichosz, Zuzanna Deckert, Monika Garnys, Izabela Grabarczyk, Wojciech Janowski, Sylwia Karasińska, Aleksandra Kujawiak, Piotr Misztela, Maria Szymańska, Karolina Walkusz, Igor Siek, Jakub Kwiatkowski, and Piotr Pęzik. 2025. [Pllum: A family of polish large language models](#). *arXiv preprint arXiv:2511.03823*.
- Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. [An open-source data contamination report for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*,

- pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.
- Jiehao Liang, Haihui Yang, Shiping Gao, and Xiaojun Quan. 2025. [Edit-wise preference optimization for grammatical error correction](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3401–3414, Abu Dhabi, UAE. 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. 2025a. [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.
- Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi, Špela Arhar Holdt, Ilze Auzina, Roberts Dargis, Elena Drakonaki, Jennifer-Carmen Frey, Isidora Glišić, Pinelopi Kikilintza, Lionel Nicolas, Mariana Romanyshyn, Alexandr Rosen, Alla Rozovskaya, Kristjan Suluste, Oleksiy Syvokon, Alexandros Tantos, Despoina-Ourania Touriki, Konstantinos Tsiotskas, Eleni Tsourilla, Vasilis Varsamopoulos, Katrin Wisniewski, Aleš Žagar, and Torsten Zesch. 2025b. [Towards better language representation in Natural Language Processing – a multilingual dataset for text-level Grammatical Error Correction](#). *International Journal of Learner Corpus Research*, 11(2):309–335.
- Marcin Miłkowski. 2010. [Developing an open-source, rule-based proofreading tool](#). *Software Practice and Experience*, 40:543–566.
- Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. [The first QALB shared task on automatic text correction for Arabic](#). In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.
- Daniel Naber. 2003. [A rule-based style and grammar checker](#). Diplomarbeit, Universität Bielefeld, Bielefeld, Germany.
- Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. [Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses](#). *Transactions of the Association for Computational Linguistics*, 7:551–566.
- 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.
- 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.
- Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. [The CoNLL-2013 shared task on grammatical error correction](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.
- Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025. [Bielik 11b v3: Multilingual large language model for european languages](#). *Preprint*, arXiv:2601.11579.
- Muhammad Reza Qorib and Hwee Tou Ng. 2022. [Grammatical error correction: Are we there yet?](#) In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2794–2800, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. [The second QALB shared task on automatic text correction for Arabic](#). In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.
- 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.
- 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.

- Yusuke Sakai, Takumi Goto, and Taro Watanabe. 2025. [IMPARA-GED: Grammatical error detection is boosting reference-free grammatical error quality estimator](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25647–25654, Vienna, Austria. Association for Computational Linguistics.
- Felix Stahlberg and Shankar Kumar. 2024. [Synthetic data generation for low-resource grammatical error correction with tagged corruption models](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 11–16, Mexico City, Mexico. Association for Computational Linguistics.
- Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. [UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language](#). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.
- Baoxin Wang, Xingyi Duan, Dayong Wu, Wanxiang Che, Zhigang Chen, and Guoping Hu. 2022. [CCTC: A cross-sentence Chinese text correction dataset for native speakers](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3331–3341, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. [Towards standardizing Korean grammatical error correction: Datasets and annotation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. [Evaluating prompting strategies for grammatical error correction based on language proficiency](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6426–6430, Torino, Italia. ELRA and ICCL.
- Yue Zhang, Bo Zhang, Haochen Jiang, Zhenghua Li, Chen Li, Fei Huang, and Min Zhang. 2023. [NaSGEC: a multi-domain Chinese grammatical error correction dataset from native speaker texts](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9935–9951, Toronto, Canada. Association for Computational Linguistics.

A Inference setup

The following system prompt was used during the evaluation:

Original: Dokonaj korekty błędów w następującym tekście zgodnie z zasadami poprawnej polszczyzny. Popraw błędy, ale nie zmieniaj stylu ani płynności wypowiedzi. Zastosuj poprawne symbole właściwe językowi polskiemu, np. cudzysłowy („ ”) oraz myślniki (–). Koniecznie zwróć tylko poprawioną wersję tekstu, bez żadnych dodatkowych wyjaśnień, komentarzy, oznaczeń zmian ani uwag! Tekst do korekty:

English translation: Correct errors in the following text in accordance with the rules of proper Polish. Correct errors, but do not alter the style or flow of the text. Use the correct punctuation marks specific to the Polish language, e.g., quotation marks (“ ”) and dashes (–). Be sure to submit only the corrected version of the text, without any additional explanations, comments, change marks, or notes! Text to be corrected:

As the user prompt, we provided the text to be corrected.

Some models produced a short introductory phrase, such as “Oto poprawiony tekst” (Eng. “The corrected text:”), so we applied simple rules to remove such fragments before evaluation.

The outputs were computed using the vLLM framework. We applied pure greedy decoding (temperature=0) without the beam search.