The MultiGEC-2025 Shared Task on Multilingual Grammatical Error Correction at NLP4CALL

Arianna Masciolini[1] **Andrew Caines**[2] **Orphée De Clercq**[3] **Joni Kruijsbergen**[3]
Murathan Kurfalı[5] **Ricardo Muñoz Sánchez**[1] **Elena Volodina**[1] **Robert Östling**[4]
[1]Språkbanken Text, SFS, University of Gothenburg, Sweden
[2]ALTA Institute & Computer Laboratory, University of Cambridge, U.K.
[3]Language and Translation Technology Team, Ghent University, Belgium
[4]Department of Linguistics, Stockholm University, Sweden
[5]RISE Research Institutes of Sweden, Stockholm, Sweden
multigec@svenska.gu.se

Abstract

This paper reports on MultiGEC-2025, the first shared task in text-level Multilingual Grammatical Error Correction. The shared task features twelve European languages (Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian) and is organized into two tracks, one for systems producing minimally corrected texts, thus preserving as much as possible of the original language use, and one dedicated to systems that prioritize fluency and idiomaticity. We introduce the task setup, data, evaluation metrics and baseline; present results obtained by the submitted systems and discuss key takeaways and ideas for future work.

1 Introduction

Following the successful 2023 shared task on Multilingual Grammatical Error Detection (Volodina et al., 2023), the Computational Second Language Acquisition (CompSLA) working group[1] presents MultiGEC-2025, a shared task in Multilingual Grammatical Error Correction.[2]

In the same vein as the previous task, the main objective of MultiGEC-2025 is to raise interest in NLP for lower-resourced languages. The task features no less than twelve European languages – namely Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian.

Contrary to traditional GEC resources, the MultiGEC dataset employed for the shared task (Masciolini et al., 2025a,b) consists of full texts. This is intended as an incentive for the development of systems able to take into account contexts larger than individual sentences.

Moreover, we distinguish a "minimal edits" and a "fluency edits" track. Minimal corrections are meant to result in texts that conform to the norms of the target language whilst preserving not only the intended meaning of the original text, but also as much as possible of its original grammar, lexis and writing style (Rudebeck and Sundberg, 2021). Fluency edits, on the other hand, may also include more extensive rephrasings aimed at producing more idiomatic language.

Evaluation is one of the biggest challenges in the organization of a shared task. The presence of the two distinct tracks mentioned above calls for using a mixture of reference-based and reference-free metrics. In addition, all automatic evaluation metrics need to be cross-lingually applicable and to work at the text level. In this paper, we propose three such evaluation metrics that were adapted for the shared task, as well as a one-shot multilingual LLM-based baseline.

An ulterior challenge is encouraging active participation, both within the short time frame of the competitive phase and beyond it, by making the dataset compiled for the shared task easily available in the long term.[3] All in all, we gathered system submissions from four different teams during the competitive phase, three of which worked with all twelve MultiGEC languages. At the time of writing, we also have received about fifty applica-

[1] spraakbanken.gu.se/en/compsla

[2] spraakbanken.gu.se/en/compsla/multigec-2025

[3] The MultiGEC data is available for download at lt3.ugent.be/resources/multigec-dataset.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina and Robert Östling. The MultiGEC-2025 Shared Task on Multilingual Grammatical Error Correction at NLP4CALL. *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2025)*. University of Tartu Library. 1–33.

Original
Hello Cristina! I am sorry to hear about you. How are you now? You got relief from your pain and how many weeks are you in the bed? I wish you will well soon.

Minimally corrected reference
Hello Cristina! I am sorry to hear about you. How are you now? You got relief from your pain**,** and how many weeks are you **in bed**? I **hope** you will **be** well soon.

Fluency-edited reference
Hello Cristina! I **was** sorry to hear about **your illness**. How are you now? **Did you get any relief** from your pain**,** and how many weeks **have** you **been in bed**? I **hope** you will **get better** soon.

Figure 1: Excerpt of a text from the Write & Improve corpus alongside a minimal correction and a fluency-edited version. Note that the latter was produced as an example for this paper and is not part of the subcorpus itself.

tions for data access (over ten of which after the end of the competition), which clearly indicates a broader interest in multilingual data and on the task of GEC.

The remainder of this paper is structured as follows. Section 2 starts with a more detailed description of the task and its two tracks, followed by an overview of the MultiGEC dataset (Section 2.1), an in-depth discussion of the three evaluation metrics selected for the task and their adaptation to our highly multilingual scenario (Section 2.2), as well as a description of our baseline system (Section 2.3). In Section 3, we briefly introduce the submitted systems and present the results they ontained in the competition. We reserve Section 4 for a discussion of the main takeaways from organizing and running this second shared task. Our conclusions, alongside some ideas for future work, are summarized in Section 5.

2 Task Setup

In modern NLP, GEC is a sequence-to-sequence task where the input is a possibly ungrammatical text, typically written by a learner, and the output a normalized or corrected version of the same text. As mentioned in the introduction, the MultiGEC-2025 shared task is organized into two tracks, each corresponding to a particular approach to correction (cf. Figure 1 for an example text corrected in both styles).

For Track 1, the goal is to rewrite texts to make them grammatically correct, i.e. adhering to the norms of the target language without altering the writing style of the original unless strictly necessary, thus following a "minimal edits" principle. Track 2, on the other hand, welcomes systems producing fluency-edited texts, i.e. corrections that are both grammatical and idiomatic.

Both tracks frame GEC as a text-level task. This was done in an attempt to stimulate the development of systems able to take into account contexts larger than traditional sentences, following a recent trend set by the widespread use of LLMs (e.g. Coyne et al. (2023); Loem et al. (2023); Fang et al. (2023); Davis et al. (2024)), which have much larger context windows than the previously dominant translation-based models for GEC (e.g. Brockett et al. (2006); Junczys-Dowmunt and Grundkiewicz (2014); Yuan et al. (2016)).

2.1 Data

We provide training, development and test data for twelve European languages (Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian) ranging from very high- to low-resourced. The data is organized into seventeen different subcorpora, all derived from pre-existing resources and compiled together into the MultiGEC dataset (Masciolini et al., 2025a,b). Table 1 provides an overview of the datasets in terms of target languages, source corpora, authorship, split sizes, amount of available correction hypothesis sets and correction styles.

As can be inferred from the table, texts come from a variety of sources. For most datasets, the authors of the texts are second language (L2) learners of the target language. This is a direct consequence of the main area of interest of the Computational SLA working group. There are, however, numerous exceptions: some of the

Language code	Subcorpus name	Source corpus	Learners	# essays (train)	# essays (dev)	# essays (test)	Ref. sets	Minimal	Fluency
cs	NatWebInf	Náplava et al. (2022)	L1	3620	1291	1256	2	✓	
	Romani		L1	3247	179	173	2	✓	
	SecLearn		L2	2057	173	177	2	✓	
	NatForm		L1	227	88	76	2	✓	
en	Write & Improve	Nicholls et al. (2024)	L2	4040	506	504	1	✓	
et	EIC	elle.tlu.ee	L2	206	26	26	3	✓	✓
	EKIL2	github.com/tlu-dt-nlp/EstGEC-L2-Corpus	L2	1202	150	151	2		✓
de	Merlin	Wisniewski et al. (2013), Boyd et al. (2014)	L2	827	103	103	1	✓	
el	GLCII	Tantos et al. (2023)	L2	1031	129	129	1	✓	
is	IceEC	Ingason et al. (2021)	L1	140	18	18	1		✓
	IceL2EC	Ingason et al. (2022)	L2	155	19	19	1		✓
it	Merlin	Wisniewski et al. (2013), Boyd et al. (2014)	L2	651	81	81	1	✓	
lv	LaVA	Daŕgis et al. (2020), Daŕgis et al. (2022)	L2	813	101	101	1	✓	
ru	RULEC-GEC	Rozovskaya and Roth (2019)	mixed	2539	1969	1535	3	✓	✓
sl	Solar-Eval	Gantar et al. (2023)	L1	10	50	49	1	✓	
sv	SweLL_gold	Volodina et al. (2019), Volodina et al. (2022)	L2	402	50	50	1	✓	
uk	UA-GEC	Syvokon et al. (2023)	mixed	1706	87	79	4	✓	✓

Table 1: Overview of the MultiGEC-2025 dataset.

Czech, Icelandic and Slovene subcorpora exclusively consist of native speaker (L1) productions – the authors being often, but not always, school children; the Russian corpus comprises essays written by both L2 and heritage speakers and the Ukrainian portion of the dataset is crowdsourced, with no information about the language background of the authors available. Additionally, proficiency levels, text genres, text lengths and subcorpus sizes vary widely across languages. On the one hand, the heterogeneity of the data means that results are not always directly comparable between different languages and subcorpora. This diversity, however, also makes it possible to compare performance between different domains and learner types.

Although most of the source corpora are error-coded, annotation is not consistent across languages. Since contemporary GEC only requires parallel texts – the original and corrected versions, often referred to as a *references* – this problem was solved by omitting all original error codes and converting all subcorpora to a simple Markdown-based format consisting of plain-text files in which alignments are indicated through essay identifiers. Notably, multiple alternative correction hypotheses are available for some of the languages (namely Czech, Estonian, Russian and Ukrainian). Corpora with multiple references are especially valuable because the reliability of reference-based metrics increases when more correction hypotheses are available (see Section 2.2).

The MultiGEC-2025 dataset is now available as a separate resource to enable future work on GEC for all languages included in the task (Masciolini et al., 2025b).[4] Alongside the data, we provide scripts to validate, parse and generate files in this format, as well as all of the evaluation scripts used in the shared task.[5] It must be noted that the current dataset release does not include gold corrections for the test splits. Evaluation of system hypotheses for test data, however, can be carried out on CodaLab[6] using one of the three evaluation metrics employed in the shared task – the GLEU score (see below).

2.2 Evaluation Metrics

As with other text generation tasks, the evaluation of GEC may be approached with reference-based or reference-free methods (Bryant et al., 2023). Reference-based evaluation metrics compare correction hypotheses to a gold standard obtained from human experts. Reference-free met-

[4] Download page: lt3.ugent.be/resources/multigec-dataset.

[5] github.com/spraakbanken/multigec-2025/tree/main/scripts

[6] codalab.lisn.upsaclay.fr/competitions/20500

rics, on the other hand, are important because they enable the evaluation of model output without relying on a single (or, at best, a few) gold-standard correction. This flexibility has become essential with the increasing popularity of LLMs in GEC, as these models are able to generate more varied but still valid corrections that may not align with human references (Östling et al., 2024). Reference-free evaluation methods were thus recently proposed as a way to estimate the quality of system output without relying on gold-standard annotations (cf. Napoles et al. (2016b); Asano et al. (2017); Choshen and Abend (2018); Yoshimura et al. (2020); Islam and Magnani (2021); Maeda et al. (2022)).

For the MultiGEC-2025 shared task, we have opted for three of the most widely used GEC evaluation metrics, each of which offers a different perspective on the quality of the proposed corrections. We use two reference-based metrics – ERRANT (Bryant et al., 2017) and GLEU (Napoles et al., 2015, 2016a) – and one reference-free metric: the Scribendi score (Islam and Magnani, 2021).

For both tracks, all system submissions were scored with these three metrics, but for each track one primary metric was chosen to obtain the final ranking. For Track 1 (minimal edits), we opted for the ERRANT-based $F_{0.5}$ score, a reference-based metric that weighs recall lower than precision, thus penalizing over-correction. For Track 2, which welcomes extensive rephrasings, we adopted the reference-free Scribendi score. We see GLEU as a useful additional metric as it is somewhere in between ERRANT and Scribendi in terms of strictness: it was designed to reward fluency rather than counting edit operations, but still relies on gold-standard corrections.

A major challenge is the need for cross-language applicability, i.e., the requirement for our scoring algorithms to be able to consistently score system output for all languages in the task. Below, we describe the evaluation metrics and steps taken to ensure that each metric can handle the twelve MultiGEC languages.

2.2.1 Reference-based metrics

The ERRANT scorer The ERRor ANnotation Toolkit (ERRANT) enables reference-based evaluation of GEC, adopting an information retrieval approach and outputting precision, recall and $F_{0.5}$ scores to represent the quality of hypothesized corrections compared to references (Bryant et al., 2017). For instance, if a system proposes four insertions of a definite article and three of them are in the correct place, then precision is 0.75; if two gold-standard insertions were missed then recall is $\frac{3}{5} = 0.6$. $F_{0.5}$ is used instead of F_1 so that precision is weighted twice as much as recall in the calculation of the F-measure, based on the reasoning that proposing incorrect corrections to learners in downstream applications is more problematic than failing to correct errors.

ERRANT was designed for English and the original implementation is publicly available.[7] It can be adapted to other languages, but in order to take advantage of its error typing and granular scoring functionality, new classification rules should be written to identify different error types (e.g. subject-verb agreement errors, word order errors, etc). Although such work has been carried out for three of the MultiGEC-2025 languages – Czech (Náplava et al., 2022), German (Boyd, 2018) and Greek (Korre et al., 2021), we had neither time nor resources to carry out this exercise for the rest of the languages and wanted to evaluate the various MultiGEC datasets in a consistent fashion. As a stop-gap measure, we added multilingual support in a rudimentary fashion for the automatic alignment of original and corrected texts, upon which holistic scoring depends. It remains to be seen in future work what impact improved adaptation of ERRANT to other languages would have on evaluation scores.

ERRANT uses spaCy[8] for part-of-speech tagging and lemmatization, which are both necessary for the alignment step. Whenever possible, fast, offline UDPipe 1 models (Straka and Straková, 2017), available through spacy-udpipe[9] were applied. In the case of Icelandic, where no such model is available, the UDPipe 2 API (Straka, 2018) was used instead.

GLEU The Generalized Language Evaluation Understanding score (GLEU) (Napoles et al., 2015, 2016a) is a reference-based metric adapted from the Bilingual Language Evaluation Understanding score (BLEU) used in MT (Papineni et al., 2002). The intuition behind GLEU is that it rewards n-grams in the model outputs that appear in the reference text but not in the original in-

[7] github.com/chrisjbryant/errant
[8] spacy.io
[9] github.com/TakeLab/spacy-udpipe

Original
Hello Cristina! I am sorry to hear about you. How are you now? You got relief from your pain and how many weeks are you in the bed? I wish you will well soon.

Gold reference
Hello Cristina! I am sorry to hear about you. How are you now? You got relief from your pain**,** and how many weeks are you **in bed**? I **hope** you will **be** well soon.

Scribendi Scoring
Original perplexity (PPL_{orig}) = 33.0
Hypothesis perplexity (PPL_{hypo}) = 25.75
Token Sort Ratio (TSR) = 0.96
Levenshtein Distance Ratio (LDR) = 0.9625
$\max(TSR, LDR) = 0.9625 > 0.8$
$PPL_{hypo} = 25.75 < PPL_{orig} = 33.0$
$\therefore Scribendi = 1$

System hypothesis
Hello Cristina! I am sorry to hear about you. How are you now? You got **some** relief from your **pain and** how many weeks are you **in bed**? I **wish** you will **be** well soon.

ERRANT Scoring
true positives (TP) = 2
false positives (FP) = 1
false negatives (FN) = 2

Precision (P) = $TP/(TP + FP) = 2/3 = 0.\dot{6}$
Recall (R) = $TP/(TP + FN) = 2/4 = 0.5$
$F_{0.5} = (1+\beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P)+R} = 2.25 \cdot \frac{0.\dot{3}}{2} = 0.1\dot{6}$

Figure 2: Worked example of ERRANT and Scribendi scoring – using the same original (top) and minimally corrected reference text (left) as in Figure 1. On the right is a created minimal correction hypothesis, in which some but not all of the reference edits have been made (note the failure to insert a comma after *pain* and to replace *hope* with *wish*). In addition, a correction is proposed which is not in the reference (insertion of *some*).

put and penalizes n-grams that are present in both the original and corrected texts but not in the reference(s). Although the GLEU score was initially proposed in Napoles et al. (2015), its implementation is presented in Napoles et al. (2016a), which offers a revised formulation that leads to a more reliable score, regardless of the number of available references. In the MultiGEC-2025 shared task, we use a later implementation by Shota Koyama, which corrects the calculation of precision.[10]

2.2.2 Reference-free metrics

Scribendi The Scribendi score (Islam and Magnani, 2021) is a reference-free metric that evaluates the quality of the corrections through a pretrained language model. The core idea is to use perplexity as a proxy for assessing both the fluency and grammaticality of the output of a GEC system. In language modeling, perplexity measures how well a model predicts a sequence of words in a given text, with lower perplexity scores indicating that the text aligns closely with the language model's predictions. Thus, low perplexity suggests that the target text closely matches the typical language usage captured by the model. Scribendi uses this alignment as an indirect measure of linguistic accuracy.

However, the perplexity score alone does not guarantee quality GEC output as perplexity does not indicate whether the intended meaning in the original text is preserved or not. As such, a GEC model that outputs only a short well-formed sentence in the target language would consistently achieve a perplexity score lower than the original text's. In order to overcome this limitation, Scribendi employs a filtering mechanism based on token ratio and Levenshtein distance and discards any corrections that orthographically deviate too significantly from the original text.

Another advantage of Scribendi is that, as long as it relies on a multilingual model, it is easily applicable to new languages, enabling cross-lingually consistent evaluation across languages which was necessary for the shared task at hand. In our preliminary experiments, we evaluated a wide range of multilingual models, with sizes ranging from 1.7 billion to 9 billion parameters, on synthetically corrupted texts across five languages. We ultimately selected Gemma 2 9B[11], which we found to be the most consistent model.

[10] github.com/shotakoyama/gleu/

[11] huggingface.co/google/gemma-2-9b-it

Scribendi assigns a score of 1, 0, or -1 to each text[12], which indicates whether the corrections lower the perplexity of the original sentence, retain or increase it. Additionally, to ensure that the hypothesis does not deviate too much from the original text, Scribendi employs orthographic similarity metrics, the Token Sort Ratio and the Levenshtein Distance Ratio[13]. If either metric falls below a threshold of 0.8, a score of -1 is assigned, indicating that the correction is too dissimilar from the original text. The overall score is calculated by adding these values across all texts. That is, a higher score indicates a greater proportion of successful corrections, with 1 meaning all corrections improve perplexity, 0 meaning none does and a negative score indicates, overall, more corrections increased the perplexity. However, as a reference-free metric, Scribendi is incapable of assessing the accuracy and quality the corrections – only their fluency and overall grammaticality. A system can achieve a perfect score even by making each sentence slightly more fluent without actually fixing all the grammatical errors.

For calculating our three chosen metrics, we convert all texts to a plain text format with one essay per line. In addition, we segment all texts with the language-agnostic `syntok` package[14] to ensure that pre-tokenized and unprocessed datasets are treated in the same way. In Figure 2 we show how evaluation works for the primary metric in each track – ERRANT for minimal correction, and Scribendi for fluency correction – using the English example from Figure 1 and an artificial 'system hypothesis' for the minimal correction track.

2.3 Baseline

The idea behind our baseline is prompting an LLM with one-shot in-context learning. As demonstrated in Davis et al. (2024), prompting LLMs is a simple but effective way to bootstrap GEC systems. However, building a single baseline for the MultiGEC dataset comes with two additional challenges. On the one hand, just as for evaluation metrics, the heterogeneity of the dataset calls for a highly multilingual model. Furthermore, both the need for reproducibility and the licensing conditions for some of the datasets impose the use of an offline, open source model.

Based on these requirements, our model of choice is the eight billion parameter, instruction-tuned version of Llama 3.1[15] (Grattafiori et al., 2024). Although prompting is only officially supported for a subset of the MultiGEC languages (English, German and Italian), this model has likely been exposed to most if not all of them during training on the continuously updated web-scraped Common Crawl dataset[16]. The latter has been shown to comprise over 170 languages, though about one third represents English data (Ortiz Suárez et al., 2019).

We use a prompt based on Davis et al. (2024), albeit with some modifications whose aim is to clearly specify the target language, distinguish between the two aforementioned correction styles, and try to prevent generation of extra text such as faux explanations:

> *You are a grammatical error correction tool. Your task is to correct the grammaticality and spelling of the input essay written by a learner of* `TARGET LANGUAGE`*.* `TASK DESCRIPTION`*. Return only the corrected text and nothing more.*

Here, `TARGET LANGUAGE` is the language of the essay at hand, while `TASK DESCRIPTION` varies based on the chosen correction style:

Minimal edits
Make the smallest possible change in order to make the essay grammatically correct. Change as few words as possible. Do not rephrase parts of the essay that are already grammatical. Do not change the meaning of the essay by adding or removing information. If the essay is already grammatically correct, you should output the original essay without changing anything.

Fluency edits
You may rephrase parts of the essay to improve fluency. Do not change the meaning of the essay by adding or removing information. If the essay is already grammatically correct and fluent, you should output the original essay without changing anything.

To further mitigate format issues in the system output, we also include a single artificial input-output pair in English, thus resulting in a one-shot-baseline.

In addition to this LLM-based system, part of the evaluation also makes use of a "dummy" zero-edit baseline. This is only relevant for establishing a lower bound for GLEU-based scoring (cf. Fig-

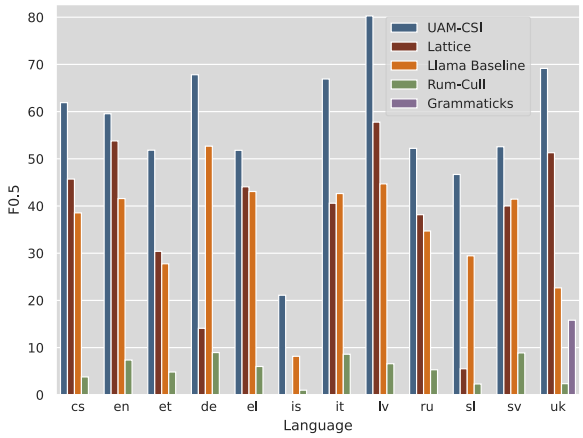
[12] The unit of analysis can be adjusted to any level, e.g. sentence, paragraph etc., but is set to full texts in our evaluation.

[13] See Islam and Magnani (2021) for details.

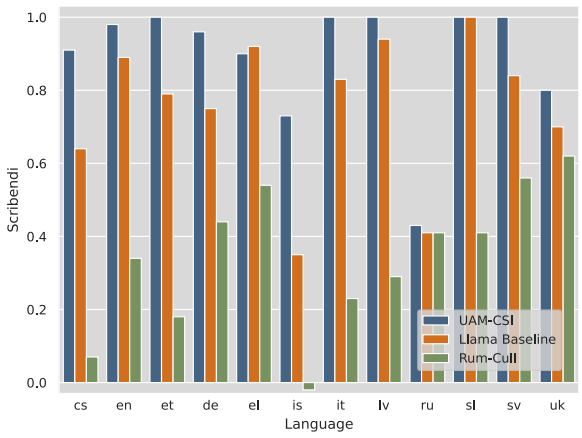
[14] `github.com/fnl/syntok`

[15] `huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

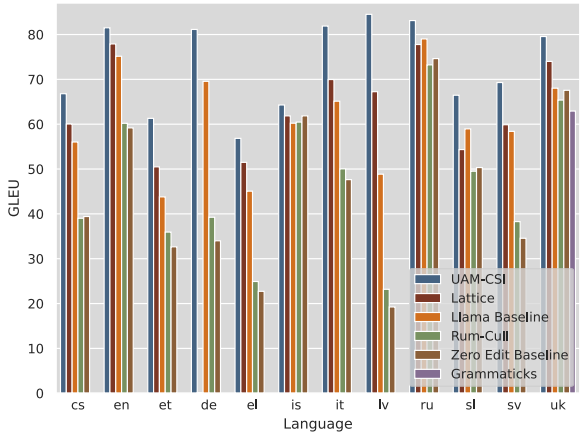
[16] `commoncrawl.org`



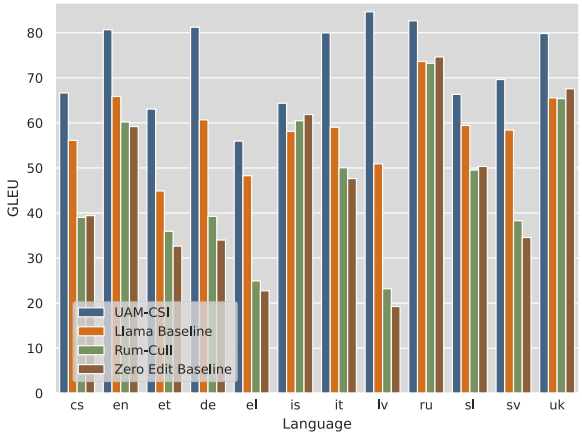
(a) $F_{0.5}$ scores (primary metric) for Track 1 submissions compared with our Llama-based baseline.



(b) Scribendi scores (primary metric) for Track 2 submissions compared with our Llama-based baseline.



(c) GLEU scores for Track 1 submissions compared with our Llama-based baseline, as well as with a zero-edit baseline.



(d) GLEU scores for Track 2 submissions compared with our Llama-based baseline, as well as with a zero-edit baseline.

Figure 3: Overview of the language-wise cross-subcorpus average scores obtained by the submitted systems for different tracks and evaluation metrics. These plots are also available in full size as part of Appendix A.

ures 3c and 3d), since $F_{0.5}$ and Scribendi scores would by definition always be equal to 0.

3 Teams, Approaches and Results

The competitive phase of the shared task ended with four submitting teams. When it comes to the "minimal edits" track (Track 1), three of them – Lattice, Rum-Cull and UAM-CSI – submitted multilingual systems addressing the GEC task for all twelve MultiGEC languages. In addition, a fourth team, Grammaticks, submitted a monolingual system for Ukrainian. Contrary to our expectations, the fluency track (Track 2) was less popular among participants and only received two submissions. Team Rum-Cull submitted the same system output to both tracks, whereas UAM-CSI used two different variants of the same systems for the two tracks.

For both tracks, the winning team is UAM-CSI. Their system, described in Staruch (2025), is the result of fine-tuning the open source LLM Gemma 2. Interestingly – but not decisively in terms of the final ranking, cf. Section 3.2 – this is the same model we selected for the Scribendi-based evaluation. The difference between the two version of this model submitted to the two tracks lies in the amount of data used for fine-tuning: for minimal edits (Track 1), only one reference file per dataset was used, whereas fluency-edited texts (Track 2) were obtained with a system fine-tuned on all available references.

Another team, Lattice, followed a similar approach for the vast majority of the languages, fine-tuning a LlaMA 3 model on MultiGEC data. The team, however, also developed an XLM-RoBERTa-based detection-correction pipeline, which they used for Slovene

given LlaMA's low performance on texts in this particular language. Both systems are described in Seminck et al. (2025).

At the time of writing, implementation details for the remaining submissions are not known to the organizers.

3.1 Automatic evaluation

As can be seen in Figure 3, team UAM-CSI is the undisputed winner of the shared task across tracks, languages and evaluation metrics, with only a handful of subcorpus-metric combinations where it is slightly outperformed by the baseline in the fluency track (cf. Appendix C).[17]

In general, our Llama-based one-shot baseline proved hard to beat. When it comes to Track 1, the winning model is the only one consistently outperforming the baseline, with the second-best system beating the latter for ten out of twelve languages in terms of GLEU scores (cf. Figure 3c) and only seven when it comes to $F_{0.5}$, the winning metric (cf. Figure 3a). As for Track 2, the UAM-CSI system scores highest in the vast majority of cases, closely followed and occasionally surpassed by the baseline (cf. Figures 3b and 3d).

Some languages appear to be especially challenging for all of the systems. In particular, all scores are exceptionally low for Icelandic, with only the winning system outperforming a zero-edit baseline in terms of GLEU (cf. Figures 3c-3d) and even a negative Scribendi score (cf. Figure 3b). At the time of writing, we did not have the opportunity to manually assess the quality of any system output for this particular language. However, we can speculate that the small size of the Icelandic subcorpora, especially when it comes to their development and test splits, results in lower scores. Furthermore, the fact that Icelandic is the only language for which only fluency edits are available might also affect the results. Finally, at least for LLM-based systems, poorer performance on this language might be due to limited exposure to the language during pre-training[18]. Russian might also suffer from the latter problem, while the surprisingly low scores that the second-best team, Lattice, obtained on German are to be attributed to the submitted system output being incomplete.

3.2 Preliminary manual evaluation

To provide more insight into the results, we performed a preliminary manual evaluation. This was done systematically on five languages, namely English, German, Italian, Russian and Swedish. While the choice of languages was mostly based on the language skills of the authors of this paper, we argue that it is also representative in a variety of senses. First of all, this selection covers all three language families represented in the dataset, Germanic (English, German and Swedish), Romance (Italian) and Slavic (Russian). Moreover, it includes a language for which several systems scored relatively high (English) one of the more challenging ones (Russian) as well as one for which we observe significant differences between teams (Italian). For each language, we selected one "challenging" case, i.e. a text whose original version greatly differs from the gold reference(s).[19]

Upon manual evaluation, both submissions by the winning team, UAM-CSI, perform generally well, especially for the three Germanic languages considered. In the vast majority of cases, the minimal correction system proposes appropriate changes, but it has a slight tendency towards under-correction. The fluency-oriented system works better overall, but sometimes leads to over-correction (e.g. for Italian). Yet, some of the more challenging issues, such as those regarding idiomatic expressions and word choice, are occasionally missed, and the system's interpretation of ambiguous or otherwise unclear sentences sometimes differs from that of the human annotators.

Team Lattice only submitted for the minimal track, where it ranked second. To the eyes of the human evaluators, the corrections proposed are reasonable on the whole, although the system occasionally introduces unnecessary or incorrect edits, such as changing plural forms to singular in Swedish and German. Furthermore, despite being submitted to the minimal track, the system

[17] Tables with the complete evaluation results for the two tracks can be found in Appendices B and C. In addition, we provide development-set results as of January 2025 (cf. Appendices D-E).

[18] Details on the exact composition of pre-training data for LLMs on a per-language basis are rarely available, including for the LLMs referred to in this paper.

[19] The manual evaluation is based on six texts with the following essay identifiers: `essay_254e63323678f4d1` (English), `1325_9000532` (Italian), `1023_0101844` and `1031_0003156` (German; in this case, two different essays were used because of the limited submission texts from team Lattice), `FL_IM_authorID-5_essayID-339_test-167` (Russian) and `G34GT1` (Swedish).

sometimes applies fluency edits (this is the case, for instance, in Italian). For Russian, on the other hand, the main issue is under-correction: the system misses over 80% of the errors that were corrected in the human reference text, mostly concerning gender and number agreement in nouns, prepositional phrases and idiomatic expressions.

Team Rum-Cull submitted the same correction hypotheses to both tracks. Across all checked languages, their system consistently applied very few changes, consisting solely of single-word replacements that often fit into the immediate context, but disregard the broader context or alter the meaning of the text. While some of the edits, especially those dealing with spelling and inflection errors, are valid, many introduce drastic semantic changes or lead to further grammaticality issues. As such, the system is currently too unreliable for end-user applications.

Overall, manual inspection confirms the viability of the evaluation metrics discussed in Section 2.2, including those that required adaptation to the highly multilingual scope of the shared task. Given the very small scale of this preliminary evaluation, it is not possible to say whether the scores are cross-lingually consistent: some of the Scribendi scores, for instance, appear suspiciously high. However, the impression we get after examining this sample of the system output is that submissions were ranked fairly. Moreover, the overall low scores for Russian correlate with our empirical observations.

Finally, in an attempt to at least partly explain the low scores registered for all systems on Icelandic data, we glanced at the relevant submissions. Without going into the merits of individual corrections, which would require more expertise in Icelandic, we notice some general trends. First of all, the top-ranking systems, i.e. the UAM-CSI submissions for the two tracks, apply very few corrections. This, however, may simply due to the fact that the texts do not require much editing, which is not unlikely given that one of the two Icelandic subcorpora consists of texts written by native speakers and the other one includes full Master's-level theses, presumably written by highly proficient L2 speakers of the language. Team Rum-Cull's submission, on the other hand, contains many single-word edits, probably overcorrecting the texts. This would explain the negative Scribendi score the team obtained for Icelandic. Finally, team Lattice's submission leaves original texts completely unchanged, thus explaining the $F_{0.5}$ score of 0.

4 Reflections and Takeaways

As mentioned, four teams submitted during the competitive phase of the shared task. Although this number was sufficient to create some competition, it was a pity that several other groups who expressed interest in participating in the task during the development phase did not eventually make any submissions on the test data. In particular, the contrast between the number of submissions and the amount of requests to access the MultiGEC dataset (approximately forty during the competitive phase of the task, increasing to fifty at the time of writing) and CodaLab registrations (twenty during the competitive phase of the task) is striking.

These numbers are evidence as to the rather strong interest in multilingual GEC, whereas the attenuation in active participation is arguably a symptom of the many demands on researchers' time and of the difficulty in developing systems for shared tasks with strict time constraints. Moreover, the task guidelines explicitly prohibited entering the data into commercial LLMs which might also have had an influence. In the following, we reflect on the issues we encountered in our role as shared task organizers and suggest ways to address them in future initiatives.

Timeline The most obvious concern is the timeline. We published our first call for participation in June 2024, a second one in September, and then released the training and development data on 21 October. This gave just over three weeks for system development and tuning until the test phase opened on 13 November. The test phase ran through to 29 November, giving participants just over a fortnight for preparing final submissions. This was an evidently tight timeline, and may have led to some teams failing to make it in time for the test phase. For comparison, the BEA 2019 shared task on GEC (Bryant et al., 2019) involved a development phase of approximately two months, followed by a test phase of just four days. Similarly, MultiGED-2023 (Volodina et al., 2023) had a 1.5-month development phase and one-week test phase. Future competitions should perhaps follow a similar approach.

Evaluation Metrics Evaluating GEC systems is not a straightforward task, with many different existing metrics and implementations to choose from (Bryant et al., 2023). It was desirable that our evaluation method should be well-grounded in the literature and previous shared tasks while being specific and suited to the datasets and languages at hand. In the case of the present shared task, we had two separate tracks calling for different evaluation strategies, data from twelve different languages, and the novelty of dealing with full texts. This introduced an array of additional constraints, which we attempted to satisfy by providing a first adaptation of three existing evaluation metrics, discussed in Section 2.2. This, however, was a time consuming process that resulted in us only disclosing the details of the evaluation procedure upon opening the competitive phase. All in all, our effort can be seen as a first step towards a cross-lingual GEC evaluation framework supporting system that work at the text level.

Benchmarking Platforms An additional layer of complexity comes from the need to fully automate the evaluation process. While this is highly desirable during the competitive phase of a shared task, where any delays affect participants, it becomes essential in cases where the competition is followed by an open phase in which system developers can participate in the task for an indefinite amount of time. Our platform of choice for this shared task, CodaLab, only fulfills this requirement for one of the metric, GLEU. On such platform, it was not possible to set up ERRANT-based or Scribendi scoring due to, respectively, installation issues and resource constraints. More generally, LLM-based evaluation poses particular challenges due to the computational resources involved. In view of future initiatives, but also to ensure that the MultiGEC dataset remains usable, we plan to investigate the available alternatives and potentially migrate the open phase of the shared task to a new platform. Furthermore, we strongly advise organizers of similar events to carefully consider the trade-off between more advanced automatic evaluation metrics and practical viability.

Baseline Since our expectation was for submitted systems to be predominantly LLM-based due to the presence of a fluency track, it was our intention to provide a strong baseline. However, our Llama-based one-shot system proved hard to beat for most of the shared task participants, and it might be the case that this has discouraged submissions of MT-based and other supervised systems, even though it is not necessarily the case that LLM-based systems will outperform supervised ones (Davis et al., 2024).

Data Access One of the main advantages of the dataset compiled for MultiGEC-2025 is that it contains data for all twelve languages in a simple uniform format. Due to licensing issues, however, data access is not entirely straightforward: while most of the training and development data can be obtained from a single repository upon agreeing to the Terms of Use, the English and Russian subcorpora require an additional sign-up and a separate download. Even more importantly, participants do not have direct access to correction hypotheses for the test splits. The reason for this is that some of the data holders of subcorpora that are not in the public domain wish to keep them private. This is a valid standpoint, as having unrestricted access to test data gives system developers the possibility to optimize for it. Moreover, by making test set references public, it can no longer be guaranteed that LLMs have not been exposed to them during pre-training. However, this does pose a problem, especially in conjunction with the evaluation issues mentioned above: participants cannot independently compute reference-based metrics on the test set and there is currently no platform able to fully automate the process. For this reason, we follow a convention emerged from previous shared tasks where data was subject to similar constraints (cf. Bryant et al. (2019)), i.e. to also report results on development data (see Appendices D and E).

5 Conclusions & Future Work

In this paper, we have provided an overview of the MultiGEC-2025 shared task. To the best of our knowledge, this is the first ever shared task on multilingual text-level GEC. We worked with twelve European languages, represented by seventeen subcorpora of texts from a variety of domains, from L2 essays to web news. These were compiled into a single dataset, MultiGEC, which provides all data in an easy-to-use uniform format. The shared task offered two tracks so that participants could choose between two different correction styles: minimal editing, the aim of which is to address grammaticality issues, or fluency editing, where the additional aim is improved idiomaticity.

Having to evaluate submissions in both styles, we opted to use three different evaluation metrics (GLEU, ERRANT scoring, and the Scribendi score) which would either reward faithfulness to the reference corrections or fluency according to a language model. Evaluation was one of the major challenges in the organization of the shared task as these metrics required adaptations to be used in this new, highly multilingual scenario. Our preliminary manual evaluation of a small sample of the results, however, suggests that our solution led to a fair ranking of the submitted systems.

Moreover, we had to deal with the technical limitations of our platform of choice, CodaLab, in terms of automation. Competitors submitted via CodaLab, but only got immediate feedback in the form of GLEU score, while the rest of the evaluation was carried out offline by the organizers. Participants in the ongoing open-phase of the shared task may still submit their corrections to CodaLab to obtain GLEU scores. In addition, we provide a program for automatic GLEU and ERRANT scoring, as well as instructions for setting up Scribendi-based evaluation locally.

Four teams participated in the official competitive phase of the shared task, and the clear winner was the UAM-CSI team with a fine-tuned Gemma 2 model for both tracks. For the most part, this model significantly outperformed our baseline. Our system proved otherwise hard to beat, especially in the 'fluency edits' track (Track 2). Moreover, scores for Icelandic and Russian were generally lower than for the other languages, which may be due to lack of exposure to these languages for LLMs during pre-training, as well as to the peculiarities of the relevant subcorpora.

We dedicated Section 4 to some reflections on the organization of this shared task. These offer insights which can be relevant for planning similar initiatives in future. The high attrition rate in participation that we observed, for instance, could be mitigated by a different timeline, increased data accessibility and further automation of the evaluation routines. While changes to the timeline are in principle easy to implement, the ease of access to the data is, in cases like ours, strongly dependent on the licensing conditions of each source corpus, something to take into account when deciding whether to prioritize the number of languages covered or the usability of the resulting resources. The technicalities of evaluation constitute an even more complex problem, calling for both further work on benchmarking platforms and in terms of development of more lightweight cross-lingually applicable metrics.

Besides these practical aspects, evaluation can be further refined. Language-specific adaptations of ERRANT would enable analysis of system performance by error type and comparisons with state-of-the-art systems could help assess where the multilingual models submitted to MultiGEC-2025 stand with respect to their language-specific counterparts. Moreover, more extensive human evaluation – which we plan to carry out for all twelve languages – would allow us to more profoundly analyze and understand the differences between systems and their continuing weaknesses, and proceed to identify ways to make further improvements to multilingual GEC.

Finally, data-wise, possible directions for future work include collecting additional data and annotations for the current MultiGEC languages so as to make the corpus more balanced and improve the robustness of reference-based evaluation, but also incorporating additional subcorpora into the MultiGEC dataset. New subcorpora could relate to L1 or L2 speakers, different age groups and a variety of genres. We would especially welcome data for languages other than the ones featuring in MultiGEC-2025, including non-European languages, and therefore would welcome contact from those with access to such datasets or planning to collect them.

All in all, despite a limited number of submissions, the shared task resulted in a new highly multilingual resource – the MultiGEC dataset, a promising novel evaluation framework for two variants of the task of GEC and at least one system with a consistently good performance across languages. The amount of requests for data access – about fifty at the time of writing – and CodaLab registrations – twenty during the competitive phase – suggest that interest in the topic is not limited to the shared task itself and encourages us to expand and improve the dataset and continue our work on automatic and manual evaluation.

Acknowledgments

Participants from Sweden have been supported by Nationella Språkbanken and Huminfra, both funded by the Swedish Research Council (2018-2024, contract 2017-00626; 2022-2024, contract

2021-00176) and their participating partner institutions, as well as the Swedish Research Council grants 2019-04129 and 2022-02909. Andrew Caines has been supported by Cambridge University Press & Assessment.

The evaluation was enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 202206725. We thank NAISS for providing computational resources under *Projects 2024/22-21 and 2023/22-1238*. Our thanks also go to Christopher Bryant for discussion around the use of ERRANT cross-linguistically.

References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adriane Boyd. 2018. Using Wikipedia Edits in Low Resource Grammatical Error Correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

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

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia. Association for Computational Linguistics.

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.

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129, New Orleans, Louisiana. 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. *arXiv preprint arXiv:2303.14342*.

Roberts Darģis, Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova, and Kristīne Pokratniece. 2022. LaVA – Latvian language learner corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 727–731, Marseille, France. European Language Resources Association.

Roberts Darģis, Ilze Auziņa, Kristīne Levāne-Petrova, and Inga Kaija. 2020. Quality focused approach to a learner corpus development. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 392–396, Marseille, France. European Language Resources Association.

Christopher Davis, Andrew Caines, Øistein 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.

Tao Fang, Jinpeng Hu, Derek F. Wong, Xiang Wan, Lidia S. Chao, and Tsung-Hui Chang. 2023. Improving Grammatical Error Correction with Multimodal Feature Integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344, Toronto, Canada. Association for Computational Linguistics.

Polona Gantar, Mija Bon, Magdalena Gapsa, and Špela Arhar Holdt. 2023. Šolar-Eval: Evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in slovstvo*, 68(4):89–108.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, and Tobias Speckbacher. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, page arXiv:2407.21783.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu, Isidora Glišić, and Dagbjört Guðmundsdóttir. 2022. The Icelandic L2 Error Corpus (IceL2EC) 1.3 (22.10). CLARIN-IS.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, and Xindan Xu. 2021. Icelandic Error Corpus (IceEC) Version 1.1. CLARIN-IS.

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.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33, Baltimore, Maryland. Association for Computational Linguistics.

Katerina Korre, Marita Chatzipanagiotou, and John Pavlopoulos. 2021. ELERRANT: Automatic Grammatical Error Type Classification for Greek. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 708–717, Held Online. INCOMA Ltd.

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, 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, Vassilis Varsamopoulos, Katrin Wisniewski, Aleš Žagar, and Torsten Zesch. 2025a. Towards better language representation in Natural Language Processing – a multilingual dataset for text-level Grammatical Error Correction. *International Journal of Learner Corpus Research*.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfali, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi-Metsoja, Špela Arhar Holdt, Ilze Auzina, Roberts Darģis, 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, Vassilis Varsamopoulos, Katrin Wisniewski, Aleš Žagar, and Torsten Zesch. 2025b. MultiGEC [dataset]. Distributed by Språkbanken Text. PID https://doi.org/10.23695/h9f5-8143.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech Grammar Error Correction with a Large and Diverse Corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

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, Matt Post, and Joel Tetreault. 2016a. GLEU without tuning. *arXiv preprint arXiv:1605.02592*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016b. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Lisa Rudebeck and Gunlög Sundberg. 2021. SweLL correction annotation guidelines. Technical report, GU-ISS Research report series, Department of Swedish, University of Gothenburg. http://hdl.handle.net/2077/69434.

Olga Seminck, Yoann Dupont, Mathieu Dehouck, Qi Wang, Noé Durandard, and Margo Novikov. 2025. Lattice @MultiGEC-2025: A spitful multilingual language error correction system using LLaMA. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, Tallin, Estonia. University of Tartu.

Ryszard Staruch. 2025. UAM-CSI at MultiGEC-2025: Parameter-efficient LLM fine-tuning for multilingual grammatical error correction. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, Tallin, Estonia. University of Tartu.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. 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.

Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.

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

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2022. *SweLL-gold [corpus]*. Språkbanken Text. Distributed via SBX/CLARIN.

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data. In *International Conference, ICT for Language Learning, 6th Edition*.

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, Ted Briscoe, and Mariano Felice. 2016. Candidate re-ranking for SMT-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 256–266, San Diego, CA. Association for Computational Linguistics.

A Overview of the official evaluation results

A.1 Track 1 (minimal edits)

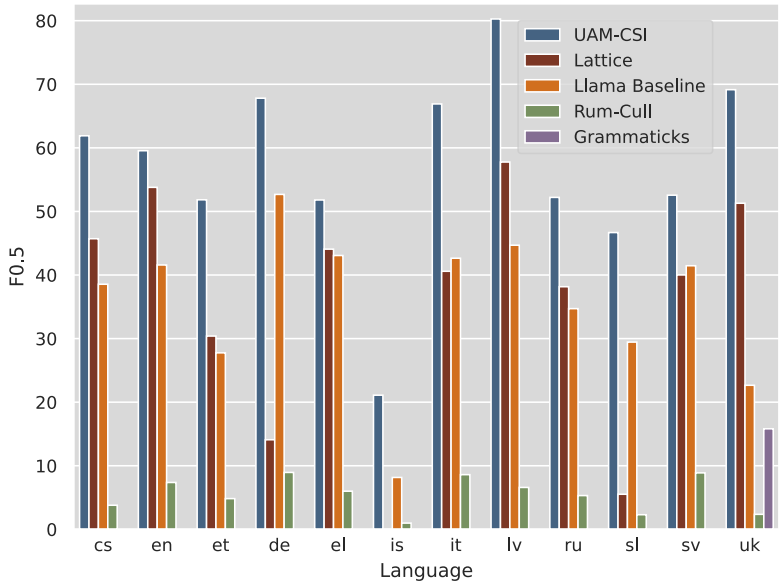


Figure A.1: Language-wise cross-subcorpus average $F_{0.5}$ scores (primary metric) for Track 1 submissions compared with our Llama-based baseline.

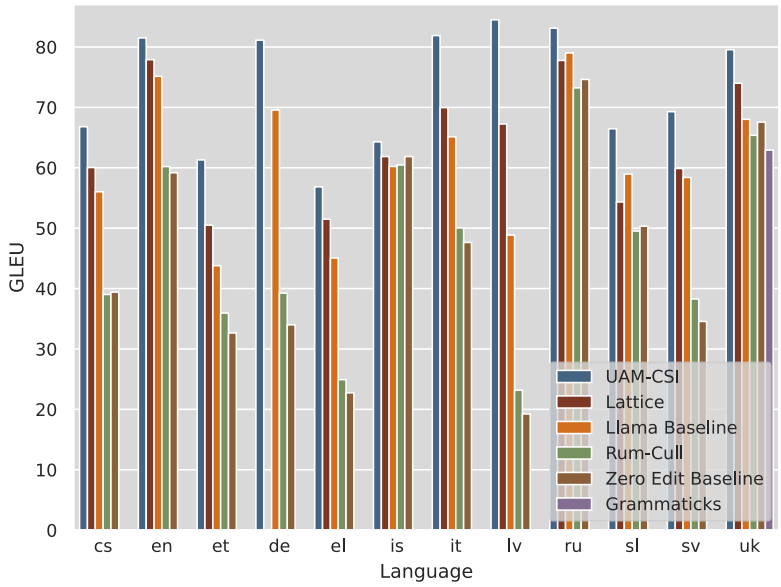


Figure A.2: Language-wise cross-subcorpus average GLEU scores for Track 1 submissions compared with our Llama-based baseline, as well as with a zero-edit baseline.

A.2 Track 2 (fluency edits)

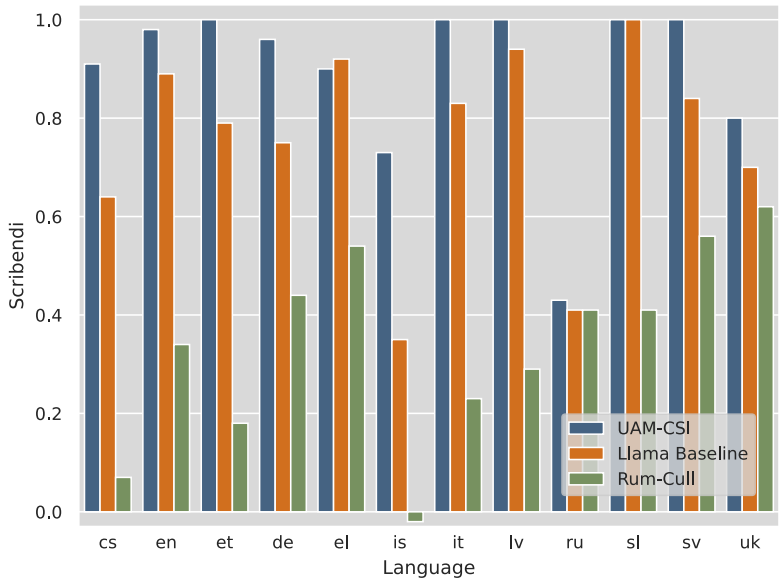


Figure A.3: Language-wise cross-subcorpus average Scribendi scores (primary metric) for Track 2 submissions compared with our Llama-based baseline.

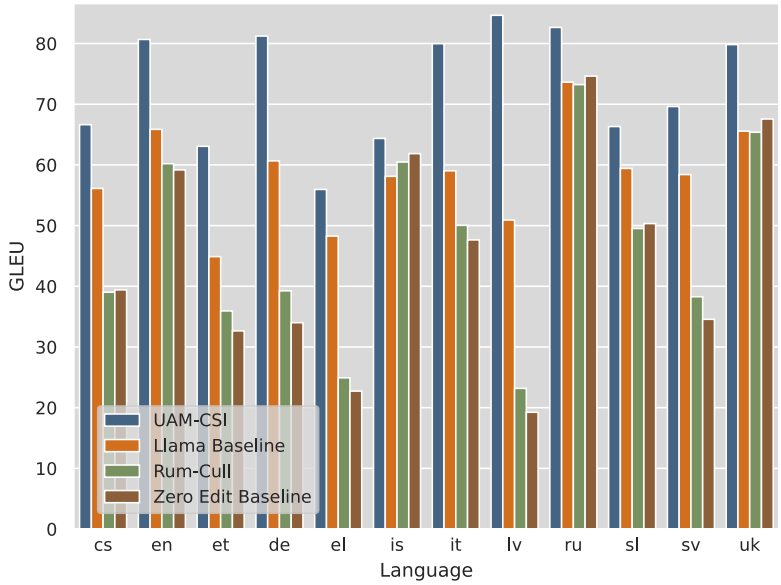


Figure A.4: Language-wise cross-subcorpus average GLEU scores for Track 2 submissions compared with our Llama-based baseline, as well as with a zero-edit baseline.

B Complete official evaluation results for Track 1 (minimal edits)

For this track, systems are ranked based on the ERRANT-based F0.5 score.

B.1 Czech

B.1.1 NatWebInf

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**69.89**	**69.81**	**63.95**	**68.55**	**0.79**
2	Lattice	65.06	56.48	55.29	56.24	0.29
3	baseline	53.91	32.89	33.06	32.93	0.74
4	Rum-Cull	40.47	3.92	1.29	2.78	0.18

B.1.2 Romani

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**60.07**	**59.94**	**50.13**	**57.68**	**0.92**
2	Lattice	53.7	48.52	38.06	45.99	0.84
3	baseline	48.35	38.52	34.52	37.65	0.82
4	Rum-Cull	26.49	6.92	1.34	3.78	0.24

B.1.3 SecLearn

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**55.81**	**62.58**	**47.23**	**58.76**	**0.98**
2	Lattice	49.95	51.69	39.26	48.61	0.94
3	baseline	45.77	50.56	34.28	46.18	0.97
4	Rum-Cull	21.92	11.17	2.77	6.96	0.34

B.1.4 NatForm

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**81.44**	**68.32**	**46.94**	**62.62**	**0.99**
2	baseline	76.08	40.41	29.0	37.46	0.92
3	Lattice	71.45	32.43	30.34	31.99	0.55
4	Rum-Cull	67.18	1.82	1.16	1.63	-0.46

B.1.5 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**66.8**	**65.16**	**52.06**	**61.9**	**0.92**
2	Lattice	60.04	47.28	40.74	45.71	0.65
3	baseline	56.03	40.59	32.72	38.55	0.86
4	Rum-Cull	39.02	5.96	1.64	3.79	0.07

B.2 English

B.2.1 Write & Improve 2024

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**81.5**	**62.24**	**50.78**	**59.55**	**0.98**
2	Lattice	77.9	58.05	41.6	53.79	0.95
3	baseline	75.15	41.59	41.55	41.58	**0.98**
4	Rum-Cull	60.2	9.63	3.8	7.37	0.34

B.3 Estonian

B.3.1 EIC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**55.76**	**54.39**	**36.23**	**49.44**	**1.0**
2	baseline	36.47	34.02	11.42	24.38	0.92
3	Lattice	44.02	22.63	23.18	22.73	0.46
4	Rum-Cull	29.06	6.83	2.06	4.66	-0.04

B.3.2 EKIL2

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**66.85**	**58.82**	**41.28**	**54.21**	**1.0**
2	Lattice	56.96	43.54	25.34	38.07	0.87
3	baseline	51.12	38.73	17.44	31.13	0.97
4	Rum-Cull	42.82	7.47	2.16	5.0	0.4

B.3.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**61.3**	**56.61**	**38.76**	**51.83**	**1.0**
2	Lattice	50.49	33.09	24.26	30.4	0.66
3	baseline	43.79	36.38	14.43	27.76	0.95
4	Rum-Cull	35.94	7.15	2.11	4.83	0.18

B.4 German

B.4.1 Merlin

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**81.13**	**68.17**	**66.43**	**67.81**	**1.0**
2	baseline	69.56	53.01	51.42	52.68	0.94
3	Lattice	0.05	30.29	4.49	14.09	-0.83
4	Rum-Cull	39.25	12.18	4.34	8.95	0.44

B.5 Greek

B.5.1 GLCII

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**56.84**	**53.79**	**45.11**	**51.8**	**0.88**
2	Lattice	51.49	45.78	38.35	44.07	0.83
3	baseline	45.07	46.95	32.39	43.07	0.97
4	Rum-Cull	24.92	12.53	1.95	6.0	0.54

B.6 Icelandic

B.6.1 IceEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**84.98**	**57.28**	**8.45**	**26.58**	**1.0**
2	baseline	80.52	9.6	5.16	8.19	0.67
3	Rum-Cull	81.18	0.85	0.43	0.71	0.22
4	Lattice	83.92	100.0	0.0	0.0	0.0

B.6.2 IceL2EC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**43.6**	**38.68**	**4.62**	**15.62**	**0.63**
2	baseline	39.93	16.88	2.65	8.14	0.26
3	Rum-Cull	39.77	2.77	0.39	1.25	-0.26
4	Lattice	39.79	100.0	0.0	0.0	0.0

B.6.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**64.29**	47.98	**6.54**	**21.1**	**0.82**
2	baseline	60.22	13.24	3.91	8.16	0.46
3	Rum-Cull	60.47	1.81	0.41	0.98	-0.02
4	Lattice	61.86	**100.0**	0.0	0.0	0.0

B.7 Italian

B.7.1 Merlin

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**81.89**	**69.04**	**59.54**	**66.91**	**0.98**
2	baseline	65.13	44.01	37.92	42.64	0.8
3	Lattice	69.96	39.9	43.65	40.59	0.85
4	Rum-Cull	50.04	11.13	4.5	8.6	0.23

B.8 Latvian

B.8.1 LaVA

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**84.5**	**80.77**	**78.32**	**80.27**	**1.0**
2	Lattice	67.25	57.8	57.61	57.77	0.9
3	baseline	48.86	47.43	36.32	44.69	**1.0**
4	Rum-Cull	23.18	10.23	2.72	6.59	0.29

B.9 Russian

B.9.1 RULEC-GEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**83.11**	**61.09**	33.01	**52.21**	**0.46**
2	Lattice	77.77	42.33	27.38	38.16	0.33
3	baseline	79.02	34.53	**35.46**	34.71	0.42
4	Rum-Cull	73.23	6.34	3.22	5.31	0.41

B.10 Slovene

B.10.1 Solar-Eval

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**66.46**	**53.89**	**30.4**	**46.68**	**1.0**
2	baseline	58.96	35.97	17.06	29.45	0.71
3	Lattice	54.34	8.67	2.25	5.52	-0.06
4	Rum-Cull	49.52	3.64	0.93	2.3	0.59

B.11 Swedish

B.11.1 SweLL_gold

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**69.29**	**54.54**	**45.88**	**52.56**	**1.0**
2	baseline	58.4	44.9	31.74	41.46	**1.0**
3	Lattice	59.88	41.49	35.02	40.01	**1.0**
4	Rum-Cull	38.28	14.02	3.6	8.88	0.56

B.12 Ukrainian

B.12.1 UA-GEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**79.55**	**74.31**	**54.11**	**69.15**	**0.89**
2	Lattice	74.0	58.55	34.28	51.29	0.1
3	baseline	68.03	26.1	14.82	22.66	0.41
4	Grammaticks	62.93	16.53	13.48	15.81	-0.1
5	Rum-Cull	65.38	3.15	1.18	2.36	0.62

C Complete official evaluation results for Track 2 (fluency edits)

For this track, systems are ranked based on the Scribendi score.

C.1 Czech

C.1.1 NatWebInf

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**70.04**	**71.05**	**64.28**	**69.58**	**0.79**
2	baseline	51.59	28.32	34.97	29.44	0.25
3	Rum-Cull	40.47	3.92	1.29	2.78	0.18

C.1.2 Romani

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**60.23**	**59.23**	**50.18**	**57.17**	**0.91**
2	baseline	48.55	33.4	33.82	33.48	0.57
3	Rum-Cull	26.49	6.92	1.34	3.78	0.24

C.1.3 SecLearn

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**55.16**	**62.21**	**46.5**	**58.27**	**0.99**
2	baseline	47.7	45.08	36.54	43.07	0.92
3	Rum-Cull	21.92	11.17	2.77	6.96	0.34

C.1.4 NatForm

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**81.07**	**68.71**	**46.82**	**62.83**	**0.95**
2	baseline	76.63	35.45	33.39	35.02	0.82
3	Rum-Cull	67.18	1.82	1.16	1.63	-0.46

C.1.5 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**66.63**	**65.3**	**51.95**	**61.96**	**0.91**
2	baseline	56.12	35.56	34.68	35.25	0.64
3	Rum-Cull	39.02	5.96	1.64	3.79	0.07

C.2 English

C.2.1 Write & Improve 2024

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**80.67**	**62.57**	**48.67**	**59.19**	**0.98**
2	baseline	65.86	24.55	40.85	26.68	0.89
3	Rum-Cull	60.2	9.63	3.8	7.37	0.34

C.3 Estonian

C.3.1 EIC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**57.89**	**56.79**	**38.6**	**51.9**	**1.0**
2	baseline	39.14	31.88	15.6	26.38	0.77
3	Rum-Cull	29.06	6.83	2.06	4.66	-0.04

C.3.2 EKIL2

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**68.23**	**56.66**	**42.86**	**53.23**	**1.0**
2	baseline	50.64	30.57	20.42	27.8	0.81
3	Rum-Cull	42.82	7.47	2.16	5.0	0.4

C.3.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**63.06**	**56.72**	**40.73**	**52.56**	**1.0**
2	baseline	44.89	31.23	18.01	27.09	0.79
3	Rum-Cull	35.94	7.15	2.11	4.83	0.18

C.4 German

C.4.1 Merlin

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**81.23**	**67.42**	**66.28**	**67.19**	**0.96**
2	baseline	60.67	44.32	50.6	45.45	0.75
3	Rum-Cull	39.25	12.18	4.34	8.95	0.44

C.5 Greek

C.5.1 GLCII

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	baseline	48.27	44.76	35.1	42.43	**0.92**
2	UAM-CSI	**55.96**	**53.62**	**44.12**	**51.4**	0.9
3	Rum-Cull	24.92	12.53	1.95	6.0	0.54

C.6 Icelandic

C.6.1 IceEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**85.09**	**61.76**	**9.03**	**28.48**	**0.72**
2	baseline	76.16	10.44	8.88	10.08	0.33
3	Rum-Cull	81.18	0.85	0.43	0.71	0.22

C.6.2 IceL2EC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**43.62**	**41.18**	4.13	**14.73**	**0.74**
2	baseline	40.08	17.86	**5.01**	11.81	0.37
3	Rum-Cull	39.77	2.77	0.39	1.25	-0.26

C.6.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**64.36**	**51.47**	**6.58**	**21.61**	**0.73**
2	baseline	58.12	14.15	6.95	10.95	0.35
3	Rum-Cull	60.47	1.81	0.41	0.98	-0.02

C.7 Italian

C.7.1 Merlin

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**79.97**	**67.45**	**56.67**	**64.98**	**1.0**
2	baseline	59.03	32.06	39.89	33.37	0.83
3	Rum-Cull	50.04	11.13	4.5	8.6	0.23

C.8 Latvian

C.8.1 LaVA

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**84.65**	**79.76**	**78.54**	**79.51**	**1.0**
2	baseline	50.92	45.57	38.92	44.07	0.94
3	Rum-Cull	23.18	10.23	2.72	6.59	0.29

C.9 Russian

C.9.1 RULEC-GEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**82.65**	**62.3**	30.94	**51.8**	0.43
2	baseline	73.63	24.02	**37.37**	25.87	0.41
3	Rum-Cull	73.23	6.34	3.22	5.31	0.41

C.10 Slovene

C.10.1 Solar-Eval

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**66.32**	**54.14**	**29.77**	**46.52**	**1.0**
2	baseline	59.42	30.84	20.58	28.04	**1.0**
3	Rum-Cull	49.52	3.64	0.93	2.3	0.41

C.11 Swedish

C.11.1 SweLL_gold

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**69.62**	**55.29**	**46.69**	**53.32**	**1.0**
2	baseline	58.41	36.62	34.0	36.06	0.84
3	Rum-Cull	38.28	14.02	3.6	8.88	0.56

C.12 Ukrainian

C.12.1 UA-GEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**79.82**	**74.65**	**55.02**	**69.68**	**0.8**
2	baseline	65.56	19.41	18.45	19.21	0.7
3	Rum-Cull	65.38	3.15	1.18	2.36	0.62

D Results on development data for Track 1 (minimal edits) as of January 2025

For this track, systems are ranked based on the ERRANT-based F0.5 score.

D.1 Czech

D.1.1 NatWebInf

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**75.64**	**67.25**	**62.7**	**66.29**	**0.76**
2	Lattice	77.64	49.31	58.35	50.88	-0.55
3	baseline	60.9	33.85	33.75	33.83	0.44

D.1.2 Romani

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**64.71**	**56.88**	**51.64**	**55.75**	**0.92**
2	Lattice	58.02	46.18	43.49	45.62	0.6
3	baseline	55.14	37.14	34.89	36.67	0.83

D.1.3 SecLearn

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**58.33**	**60.66**	**49.67**	**58.09**	**0.99**
2	Lattice	52.63	48.79	41.49	47.13	0.94
3	baseline	46.35	46.76	34.94	43.79	0.97

D.1.4 NatForm

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**84.41**	**63.32**	**49.79**	**60.05**	**0.95**
2	baseline	81.16	44.56	36.32	42.62	0.8
3	Lattice	79.52	40.58	36.14	39.61	-0.02

D.1.5 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**70.77**	**62.03**	**53.45**	**60.05**	**0.91**
2	Lattice	66.95	46.21	44.87	45.81	0.24
3	baseline	60.89	40.58	34.98	39.23	0.76

D.2 English

D.2.1 Write & Improve 2024

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**82.6**	**62.62**	**49.86**	**59.57**	**0.99**
2	Lattice	79.44	54.35	40.2	50.78	0.36
3	baseline	76.43	39.25	42.24	39.82	0.98

D.3 Estonian

D.3.1 EIC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**55.5**	**50.52**	**33.84**	**45.98**	**1.0**
2	Lattice	49.54	32.36	27.51	31.26	0.31
3	baseline	36.01	32.08	10.8	23.01	0.92

D.3.2 EKIL2

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**66.89**	**52.45**	**37.08**	**48.43**	**0.98**
2	Lattice	57.71	36.48	21.86	32.17	0.35
3	baseline	54.48	34.81	15.15	27.64	0.84

D.3.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**61.19**	**51.48**	**35.46**	**47.2**	**0.99**
2	Lattice	53.63	34.42	24.69	31.71	0.33
3	baseline	45.25	33.45	12.98	25.32	0.88

D.4 German

D.4.1 Merlin

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**80.31**	**67.23**	**63.76**	**66.51**	**0.94**
2	baseline	69.16	51.89	50.55	51.61	0.9
3	Lattice	0.0	31.8	3.35	11.79	-0.92

D.5 Greek

D.5.1 GLCII

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**57.57**	**53.56**	**44.7**	**51.52**	0.84
2	Lattice	54.68	48.47	40.17	46.55	0.74
3	baseline	47.68	49.83	33.17	45.29	**0.9**

D.6 Icelandic

D.6.1 IceEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**88.62**	**34.57**	6.53	**18.59**	**0.5**
2	baseline	85.3	9.84	**7.23**	9.18	0.22
3	Lattice	0.88	0.0	0.0	0.0	-1.0

D.6.2 IceL2EC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	baseline	45.05	**26.78**	**6.18**	**16.06**	0.16
2	UAM-CSI	**48.19**	22.87	3.99	11.75	**0.89**
3	Lattice	2.52	0.4	1.25	0.46	-0.89

D.6.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**68.4**	**28.72**	5.26	**15.17**	**0.7**
2	baseline	65.17	18.31	**6.71**	12.62	0.19
3	Lattice	1.7	0.2	0.62	0.23	-0.95

D.7 Italian

D.7.1 Merlin

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**80.27**	**68.3**	**60.9**	**66.68**	**0.98**
2	Lattice	77.15	53.78	58.11	54.6	0.58
3	baseline	66.5	50.66	43.83	49.13	0.85

D.8 Latvian

D.8.1 LaVA

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**83.89**	**81.32**	**78.62**	**80.76**	**1.0**
2	Lattice	69.09	61.73	61.33	61.65	0.94
3	baseline	47.3	48.44	38.14	45.96	0.98

D.9 Russian

D.9.1 RULEC-GEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**84.68**	**52.98**	35.63	**48.28**	0.37
2	baseline	77.82	24.84	36.54	26.54	**0.43**
3	Lattice	79.35	21.03	**36.86**	23.01	-0.86

D.10 Slovene

D.10.1 Solar-Eval

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**67.22**	**57.4**	**32.89**	**49.95**	**1.0**
2	baseline	59.55	37.2	18.86	31.14	**1.0**
3	Lattice	29.69	14.17	6.77	11.63	-0.12

D.11 Swedish

D.11.1 SweLL_gold

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**72.01**	**58.83**	**50.61**	**56.98**	**1.0**
2	baseline	56.48	48.46	30.79	43.47	0.92
3	Lattice	59.95	45.02	35.33	42.68	0.88

D.12 Ukrainian

D.12.1 UA-GEC

Rank	Team	GLEU	Precision	Recall	**F0.5**	Scribendi
1	UAM-CSI	**77.62**	**70.34**	**49.31**	**64.81**	**0.9**
2	Lattice	66.01	33.96	27.53	32.45	-0.33
3	baseline	67.29	25.15	16.61	22.81	0.64

E Results on development data for Track 2 (fluency edits) as of January 2025

For this track, systems are ranked based on the Scribendi score.

E.1 Czech

E.1.1 NatWebInf

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**76.34**	**68.68**	**64.25**	**67.74**	**0.76**
2	baseline	56.9	26.37	32.49	27.4	0.39

E.1.2 Romani

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**64.68**	**58.34**	**51.58**	**56.85**	**0.92**
2	baseline	50.87	32.12	33.71	32.43	0.7

E.1.3 SecLearn

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**58.11**	**60.58**	**49.55**	**58.0**	**0.98**
2	baseline	47.01	43.0	36.28	41.46	0.94

E.1.4 NatForm

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**84.55**	**65.47**	**50.7**	**61.86**	**0.95**
2	baseline	78.93	31.78	36.73	32.66	0.89

E.1.5 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**70.92**	**63.27**	**54.02**	**61.11**	**0.9**
2	baseline	58.43	33.32	34.8	33.49	0.73

E.2 English

E.2.1 Write & Improve 2024

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**81.98**	**63.68**	**47.61**	**59.65**	**0.98**
2	baseline	66.34	21.76	40.96	24.01	0.91

E.3 Estonian

E.3.1 EKIL2

E.3.2 EIC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**58.19**	**53.13**	**38.62**	**49.42**	**1.0**
2	baseline	37.04	32.37	14.13	25.73	0.77

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**67.91**	**52.61**	**38.92**	**49.15**	**0.99**
2	baseline	52.5	27.9	17.93	25.11	0.84

E.3.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**63.05**	**52.87**	**38.77**	**49.28**	**0.99**
2	baseline	44.77	30.13	16.03	25.42	0.8

E.4 German

E.4.1 Merlin

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**81.05**	**67.53**	**64.36**	**66.87**	**0.96**
2	baseline	65.22	43.85	49.66	44.9	0.84

E.5 Greek

E.5.1 GLCII

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	baseline	50.39	47.79	36.09	44.88	**0.91**
2	UAM-CSI	**57.98**	**53.88**	**44.7**	**51.76**	0.84

E.6 Icelandic

E.6.1 IceEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**88.29**	**28.12**	4.2	**13.14**	**0.61**
2	baseline	79.56	7.6	**8.62**	7.78	0.33

E.6.2 IceL2EC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**48.62**	21.28	3.64	10.8	**0.89**
2	baseline	45.71	**22.15**	**8.13**	**16.47**	0.58

E.6.3 Cross-subcorpus average

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**68.46**	**24.7**	3.92	11.97	**0.75**
2	baseline	62.64	14.88	**8.38**	**12.12**	0.46

E.7 Italian

E.7.1 Merlin

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**80.04**	**70.3**	**60.11**	**68.0**	**0.98**
2	baseline	56.85	35.03	41.54	36.16	0.85

E.8 Latvian

E.8.1 LaVA

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**83.32**	**81.29**	**78.95**	**80.81**	**0.98**
2	baseline	47.97	44.89	38.78	43.52	0.94

E.9 Russian

E.9.1 RULEC-GEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	baseline	72.64	18.99	**39.04**	21.17	**0.47**
2	UAM-CSI	**83.95**	**53.71**	32.03	**47.3**	0.34

E.10 Slovene

E.10.1 Solar-Eval

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**66.99**	**57.16**	**32.49**	**49.63**	**1.0**
2	baseline	59.84	30.57	22.56	28.54	**1.0**

E.11 Swedish

E.11.1 SweLL_gold

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**70.38**	**57.81**	**48.29**	**55.62**	**1.0**
2	baseline	60.05	39.27	34.7	38.26	1.0

E.12 Ukrainian

E.12.1 UA-GEC

Rank	Team	GLEU	Precision	Recall	F0.5	**Scribendi**
1	UAM-CSI	**77.54**	**69.12**	**50.95**	**64.52**	**0.9**
2	baseline	64.65	18.88	20.66	19.21	0.77