

Introducing OmniGEC: A Silver Multilingual Dataset for Grammatical Error Correction

**Roman Kovalchuk^{1,2} ,
Mariana Romanyshyn³, and
Petro Ivaniuk²**

¹UCU, ²SoftServe, ³Grammarly

MSci PROGRAM in DATA SCIENCE



Lviv – Ukraine | 31th of July, 2025

Acknowledgments

- We express gratitude to **APPS UCU** for computing resources and OpenAI API access.
- We express our gratitude to the volunteers, students, and lecturers of the National Technical University “Kharkiv Polytechnic Institute” who joined and promoted our annotation project.
- We thank **Oleksandr Skurzhanskyi**, Applied Research Scientist at Grammarly, and other reviewers of this study for their invaluable input.

Outline

1. Motivation
2. Related Work
3. Problem Setting
4. Approach:
 - New GEC Datasets – OmniGEC
 - Dataset Quality Evaluation
 - LLM Fine-Tuning Experiments
4. Outcomes
5. Limitations and Future Work

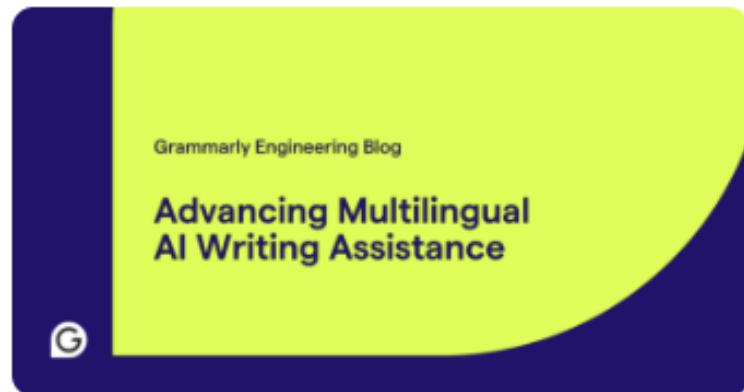
Motivation

Multilinguality is an increasingly popular topic not only in Grammatical Error Correction (GEC) but also in Natural Language Processing (NLP)

OpenAI's GPT-4o, GPT-4o-mini, o1, o3-mini, and GPT-4.5 models are marketed as more capable models not only because of their bare power in English but also because of **their ability to process low- and mid-resource languages.**

Motivation: GEC Research Papers

Multilinguality is an apparent trend in new GEC papers, following the same trend in the LLMs



Advancing AI-Powered Intelligent Writing Assistance across Multiple Languages

The Strategic Research team at Grammarly is constantly exploring how LLMs can contribute to our mission of improving lives by...

December 6, 2024

Grammarly Engineering Blog:

<https://www.grammarly.com/blog/engineering/advancing-intelligent-writing/>

Motivation: GEC Competitions

NLP4CALL Workshop hosted two Multilingual GEC competitions in the last two years

Computational
SLA

MultiGEC dataset

MultiGEC-2025

Mini-workshop on GEC 2024

MultiGEC-2023

MultiGEC-2025

View this page on [the dedicated MultiGEC website](#).

MultiEC-2025

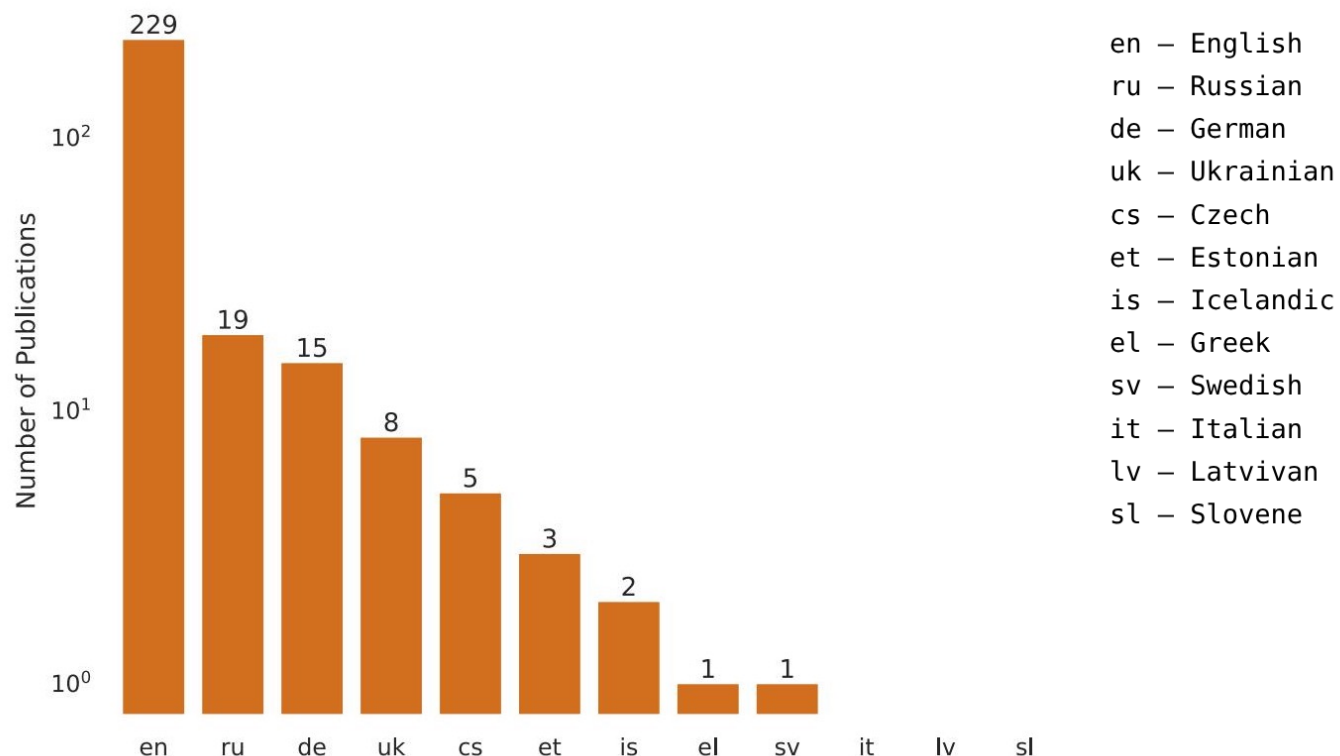
Call for participation

The [Computational SLA](#) working group invites you to participate in the shared task on text-level Multilingual Grammatical Error Correction, **MultiGEC**, covering 12 languages: Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian (see also the [call for participation on the ACL portal](#)).

<https://spraakbanken.gu.se/en/projects/computational-sla>

Motivation: GEC and Multilinguality

The number of relevant of papers is limited, even with the recent MultiGEC-25 and MultiGED-23 competitions



[The MultiGEC-2025 Shared Task on Multilingual Grammatical Error Correction at NLP4CALL](#)

Related Work

1. English (Monolingual) GEC:

- English GEC methods like high-diversity and ranking can be adapted to other languages.
- Other methods, like GECTOR, the ranking model, are specifically designed to work with English, making the transition problematic.

2. Multilingual GEC:

- An underresearched area with very few published papers.
- Approaches like synthetic data generation, mentioned in the UNLP 2023 Shared Task on Grammatical Error Correction for Ukrainian, can be adapted to our case.

3. Multilingual LLMs:

- Open-source models: **AYA-Expansive (8B)** and **Gemma (12B)**
- Proprietary models: **GPT-4o-mini** and **o1-preview**

Related Work: Pre-Trained LLMs

| Model Name | Number of Languages | Comment |
|------------------|---|--|
| AYA-ExpansE (8B) | 23 languages, including Ukrainian | More compact and modern model than AYA-101, which performed well in GEC. |
| Gemma-3 (12B) | ? – the authors do not explicitly state which languages the model targets, other than "out-of-box" support for 35 languages and pre-trained support for over 140 languages. | Highly capable multilingual model, including GEC tasks; improved performance from Gemma-2. |

Related Work: Datasets

| Dataset Name | # Examples | Number of Languages | Comment |
|---------------|------------|--|---|
| MultiGEC-2025 | 7,700 | 12 – Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Slovene, russian, Swedish, and Ukrainian | High-quality compilation of monolingual GEC datasets from human-annotators and learners |
| cLang-8 | 100,000 | 1 – English | High-quality and high-quantity machine-refined dataset |
| CoNLL-2014 | 2,000,000 | 1 – English | High-quantity of texts written by non-native speakers, annotated by humans |
| BEA-2019 | 800,000 | 1 – English | High-quality and high-quantity with a more essay variations |

***In bold – Target languages for our work.**

Related Work: GEC Competitions

| Competition | Number of Languages | Comment |
|---------------------------|---|--|
| MultiGEC-2025 | 12 – Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Slovene, russian, Swedish, and Ukrainian | The top competitors limited themselves to fine-tuning the dataset provided by the authors without per-language IT templates – simple workflows resulted in low scores. |
| MultiGED-2023 | 5 – Czech, English, German, Italian and Swedish. | Same data as in MultiGEC-2025; the methods are outdated and do not involve LLMs. |
| The UNLP-2023 Shared Task | 1 – Ukrainian. | Although the competition covers only 1 target language, the methods laid out there are easily adaptable for multilingual setting |

Related Work: Metrics

| Metric Name | Implementation | Description |
|----------------------|-----------------------------|---|
| Edit Distance | Heuristic | Fewest edits to match the reference text. |
| Character Error Rate | Heuristic | Edits required per 100 characters of reference text. |
| $F_{0.5}$ | Heuristic | Averages precision and recall from corrected errors. Gives twice as much weight to precision as to recall. |
| ERRANT | Heuristic | Analyzes pairs of original and corrected sentences and classifies the edits made between them according to a rule-based error-type framework. |
| GLEU | Heuristic (N-gram based) | N-gram-based metric used to calculate precision, recall, and $F_{0.5}$. |

Problem Setting: Gap Identification

1. Data Collection:

- High-quality GEC data is scarce
- Multilingual data is not uniform

2. Data Composition:

- Lack of data quality versus data quantity ablation studies
- Lack of cross-language ablation studies

3. Low Solution Performance:

- The results of MultiGEC-2025 do not beat SOTA for individual languages

Problem Setting: Addressing the Gaps

1. **Data Collection** → **Publish New GEC Datasets:**

- Collect uniform multilingual data
- Automatically collect silver GEC annotations
- Assess data quality with manual and automatic evaluation

2. **Data Composition** → **Conduct Ablation Studies:**

- Study impact on a per-language basis
- Study impact of data quality vs quantity on a per-dataset basis

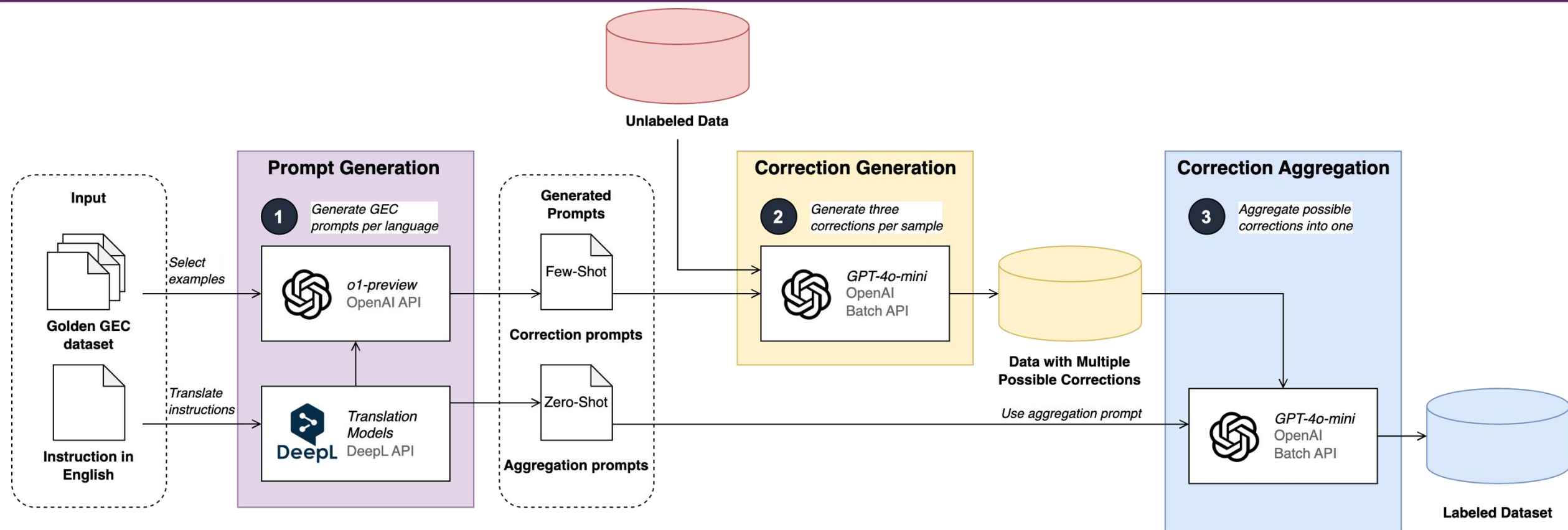
3. **Low Solution Performance** → **Utilize Techniques from English GEC:**

- Apply fine-tuning to open-source large language models of different sizes
- Collect preference data for Ukrainian as a case study for DPO fine-tuning
- Adapt methods and techniques that worked well in English GEC

New Silver GEC Datasets: OmniGEC

| Corpus | # Tokens (total) | # Samples (per-language) | Languages | Annotation Source | Data Source |
|--------------------|------------------|--------------------------|--------------|------------------------|-----------------------------------|
| WikiEdits-MultiGEC | 1.3M | ~1.6k | 11 languages | Wikipedia contributors | Wikipedia |
| Reddit-MultiGEC | 13M | ~15.4k | 11 languages | GPT-generated | Reddit |
| UberText-GEC | 22M | 200k | Ukrainian | GPT-generated | UberText 2.0 Social Media Corpora |

Synthetic Data Generation



- ❑ Leverage LLMs, human examples, and translation models to come up with LLM prompts
- ❑ Synthetic Data Generation allows to obtain relatively cheap and quality data for training

Quality Evaluation

We performed automated and human evaluation to measure the dataset quality on a subset of 1,500 samples per corpus in Ukrainian.

| Method | Pros | Cons | Reference |
|----------------------|-----------------------------|---|--|
| Automated Evaluation | Scalable, cheap and uniform | Limited | Pravopysnyk, Spivavtor, and LanguageTool |
| Human Evaluation | Accurate and representative | Low-scale, expensive, and requires language expertise | Annotator-assigned grade |

Automated Quality Evaluation

| Corpus | Precision | Recall | F _{0.5} | GLEU | Edit Distance | Character Edit Rate |
|--------------------|--------------|--------------|------------------|--------------|---------------|---------------------|
| Reddit-MultiGEC | 17.92 | 59.51 | 20.84 | 46.89 | 36.87 | 18.20 |
| UberText-GEC | 16.83 | 56.81 | 19.59 | 63.45 | 23.51 | 10.98 |
| WikiEdits-MultiGEC | 13.30 | 26.03 | 14.74 | 71.35 | 18.21 | 4.79 |

Multi-reference automated evaluation metrics across corpora with ERRANT (precision, recall, and F0.5), Levenshtein distance (error distance), Character Error Rate (normalized error distance), and GLEU

Human Evaluation

Annotation Project:

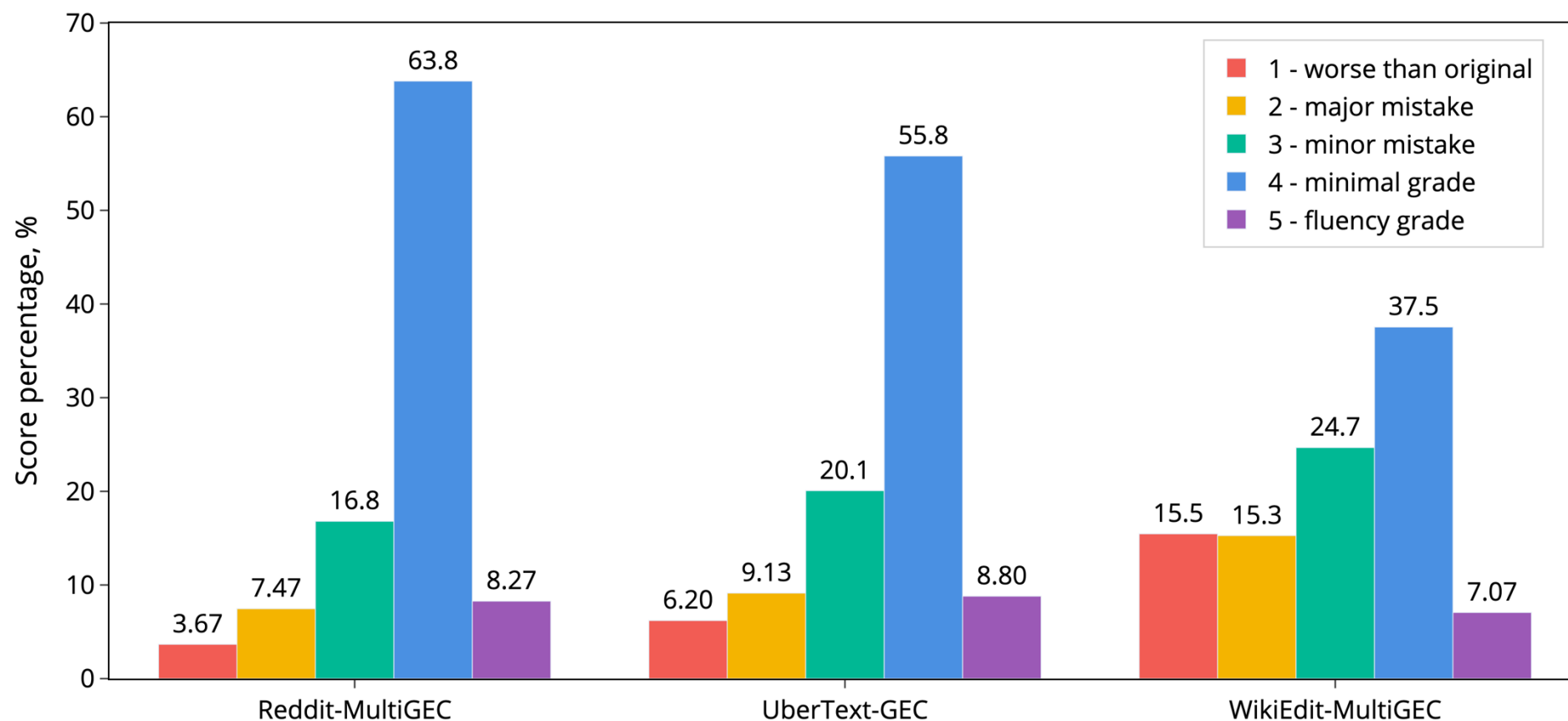
- 15 annotators, native speakers of Ukrainian with a major in linguistics
- 3x1,500 data samples
- 100 double-annotated samples for calibration

Quality Grades:

1. New errors are introduced, or the meaning has changed
2. Major errors are present, unchanged in meaning
3. Significant improvement, but minor errors remain
4. Grammatically correct, a.k.a. the “minimal” grade
5. Grammatically correct and fluent, a.k.a. the "fluency" grade

Human Evaluation: Results

More than 70% of the synthetic samples are scored as “4 – minimal grade” or “5 – fluency grade”



Data Example: Fluency Grade

Боюсь стоматолога більше чим шахеда страх стоматології відчувала лише одного разу, коли йшла до нього перший раз. Насправді мені пощастило з дантистом, і від самого дитинства він не міняється. З цього я напевно можу дати пораду, оберіть собі стоматолога котрому будете довіряти, тобто приємну для вас людину. Не менш важливою для мене є клініка. Оберіть місце яке не буде описувати дитячі страхи або фільми жахів. Комфорт найважливіше

Я боюся стоматолога більше, ніж шахеда. Страх стоматології відчула лише один раз, коли ішла до нього вперше. Насправді мені пощастило з дантистом, і з дитинства він не змінюється. Тому я можу дати пораду: оберіть стоматолога, якому будете довіряти, тобто приємну для вас людину. Не менш важливою для мене є клініка. Оберіть місце яке не нагадує про дитячі страхи або фільми жахів. Комфорт — найважливіше.



Data Example: Worse Than Original

The Rasmus співають пісню Stefania Kalush Orchestra The Rasmus цього року представляють на Євробаченні Фінляндію. І за лаштунками пісенного конкурсу зустрілися із українськими музикантами. Kalush Orchestra представить Україну під час першого півфіналу вже 10 травня.

Група The Rasmus виконує пісню Stefania на Євробаченні, представляючи Фінляндію. І за лаштунками конкурсу вони зустрілися з українськими музикантами. Kalush Orchestra представить Україну під час першого півфіналу 10 травня.



LLM Fine-Tuning Setup

1. Use zero-shot prompting with Aya-23-8B to develop the instruction templates.
2. Instruction-tune Aya-expanse (8B) and Gemma-3 (12B) with those templates in three incremental setups:
 1. MultiGEC
 2. MultiGEC+WikiEdits
 3. MultiGEC+WikiEdits+Reddit.
3. Measure how additional, language-aligned corpora translate into performance in multilingual GEC with GLEU.

LLM Fine-Tuning Results

| Model | GLEU ^{mean} _{minimal} | GLEU ^{mean} _{fluency} | GLEU ^{Ukrainian} _{minimal} | GLEU ^{Ukrainian} _{fluency} | GLEU ^{Estonian} _{minimal} | GLEU ^{Latvian} _{minimal} |
|-----------------------------|---|---|--|--|---|--|
| Our Results | | | | | | |
| Aya-23-Expansive-8B | | | | | | |
| <i>MultiGEC</i> | 64.52 | 48.37 | 77.28 | 76.51 | 33.27 | 72.29 |
| <i>MultiGEC+Wiki</i> | 65.16 | 48.37 | 77.05 | 77.10 | 38.07 | 73.04 |
| <i>MultiGEC+Wiki+Reddit</i> | 65.43 | 49.80 | 76.41 | 75.82 | 41.52 | 71.71 |
| Gemma-3-12B-IT | | | | | | |
| <i>MultiGEC</i> | 61.43 | 48.66 | 74.25 | 74.22 | 54.74 | 54.05 |
| <i>MultiGEC+Wiki</i> | 67.02 | 52.34 | 75.17 | 71.88 | 55.12 | 81.54 |
| <i>MultiGEC+Wiki+Reddit</i> | 66.42 | 49.20 | 75.11 | 74.83 | 57.54 | 80.19 |
| MultiGEC-25 | | | | | | |
| LLaMA-3-8B | | | | | | |
| <i>MultiGEC</i> | 56.85 | - | 74.00 | - | 44.02 | 67.25 |

The comparison of paragraph-based GEC models fine-tuned on the MultiGEC-25 and OmniGEC datasets across all languages and specifically for Ukrainian, Estonian (minimal), and Latvian.

Our models are compared against the MultiGEC-2025 paragraph-level GEC model

Deliverables: New GEC Datasets

1. New silver multilingual GEC dataset collection published on HuggingFace – [OmniGEC](#):
 - Ubertext-GEC (Ukrainian)
 - Reddit-MultiGEC (Multilingual)
 - WikiEdits-MultiGEC (Multilingual)
2. New preference data (annotations) for the Ukrainian part of OmniGEC are published on HuggingFace as part of [OmniGEC](#).
3. The code repository for data collection and processing is openly available to facilitate contributions and updates – [OmniGEC-Data](#).

Deliverables: Models and Paper

1. New SOTA paragraph-editing multilingual GEC models published on [HuggingFace](#):
 1. OmniGEC *Minimal/Fluency* (8B) – based on AYA-Expanse (8B)
 2. OmniGEC *Minimal/Fluency* (12B) – based on Gemma-3 (12B)
2. The code repository for instruction templates and LoRA fine-tuning is openly available for reproducibility – [OmniGEC-Models](#).
3. A supplementary research paper accepted to UNLP 2025.

Conclusions: Addressed Gaps

1. **Data Collection Challenges** → **Published New GEC Datasets:**

- Collected uniform multilingual data
- Automatically collected GEC annotations
- Assessed data quality with manual and automatic evaluation

3. **Low Solution Performance** → **Utilized Techniques from English GEC:**

- Applied fine-tuning to open-source large language models of different sizes
- Collected preference data for Ukrainian as a case study for Direct-Preference Optimization (DPO) fine-tuning
- Adapted methods and techniques that worked well in English GEC

Conclusions: Gaps To Be Addressed

2. **Data Composition Challenges** → **Conduct Ablation Studies:**


- Study impact on a per-language basis
- Study impact of data quality vs quantity on a per-dataset basis

Limitations

1. OmniGEC covers only eleven languages, leaving aside the vast linguistic diversity.
2. Human annotation was collected only for the Ukrainian language, limiting assessment and DPO tuning for other languages.
3. We used proprietary models for synthetic data generation, which may impact the reproducibility.
4. Due to time and resource limitations, Gemma-3-12B-IT was only trained for one epoch on the *MultiGEC+WikiEdits+Reddit* scenario, limiting our research to two open-source multilingual LLMs.

Future Work

1. Conduct human evaluation for the new datasets in more target languages.
2. DPO Fine-Tune model for more target languages.
3. Explore Full Fine-Tuning.



Thanks for your Attention!

Happy to answer your questions



fb.com/csatucu



[@ucu_apps](https://www.instagram.com/ucu_apps)

apps@ucu.edu.ua



apps.ucu.edu.ua



Faculty of Applied Sciences
Ukrainian Catholic University
Kozelnytska st. 2a, Lviv,
79076, Ukraine