# LLMs in alliance with Edit-based Models: Advancing In-Context Learning for Grammatical Error Correction by Specific Example Selection

Alexey Sorokin[1,2], Regina Nasyrova[1]

[1]MSU AI Institute
[2]Yandex

BEA 2025 (ACL Workshop)
August, 1st, 2025

# Our contribution

- We create a new rule-oriented corpus for Russian GEC (LORuGEC).
- We show that fewshot learning behaves comparably to finetuning on this corpus.
- We demonstrate that the quality of fewshot learning heavily depends from in-context examples' selection.
- We show the utility of GECTOR-like models for fewshot examples selection.

# Motivation

- (Russian) GEC is too easy for modern LLMs.
- Results on GERA corpus (Sorokin, Nasyrova, 2024)

| Model | Method | P | R | F0.5 |
|---|---|---|---|---|
| ruGPT2-large | finetune | 73.4 | 23.4 | 51.4 |
| ruGPT2-large+reranking | finetune | 78.4 | 44.4 | 68.0 |
| rule_generator+reranking | finetune | 86.1 | 42.9 | 71.6 |
| Qwen2.5-7B | finetune | 74.3 | 48.2 | 67.1 |
| YandexGPT-5 8B | zero-shot | 70.8 | 53.3 | 66.5 |
| YandexGPT-5 8B | finetune | 78.0 | 59.0 | 73.3 |

- Simple LLM finetuning outperforms earlier SOTA methods.
- But is the problem challenging enough?

# Russian punctuation

- Russian punctuation is rather complex:
- Normally, a comma is inserted between two coordinative clauses:

  Солнце зашло, и наступила темнота

  The sun set, and it became dark
- But not if they have a common dependent word:

  В 8 часов солнце зашло и наступила темнота

  At 8 o'clock the sun set and it became dark
- Such 'school exam' cases are underrepresented in existing corpora.
  - L2 learners do not use complex constructions.
  - Native learners prefer simpler constructions to reduce risk of errors.
- What if we collect them intentionally?

# Corpus collection

- Stage 1: design a list of complex rules.
  - 10 seed rules were selected by the authors of the papers.
  - The list was further extended by 3 linguistics students.
  - They were instructed to consult official sources such as spelling and punctuation handbooks, educational websites and books, academic dictionaries etc.
- Stage 2: collect sentences for each rule:
  - write up to 10 sentences regulated by a particular rule.
  - Avoid direct citing from fiction and reference books.
  - Pass the sentences through a LLM (Yandex GPT-3) in order to prefer complex sentences.
- Stage 3: introduce a mistake violating the rule in question:
  - If there are multiple possible mistake types, use all of them.
  - For most source sentences, a single mistake was introduced.
- Postprocessing: the rules and sampled examples were verified by one of the paper authors.

## Corpus characteristics

- The rules cover all aspects of Russian grammar:

| Subset | Count |
|---|---|
| Grammar | 4 |
| Punctuation | 17 |
| Semantics | 2 |
| Spelling (single-word) | 11 |
| Spelling (multiword) | 14 |
| Total | 48 |

- The corpus is created for diagnostics and OOD evaluation of GEC models. Thus we don't collect a large training set.

| **Sample** | **Sentences** | **Corr. source** | **Tokens** |
|---|---|---|---|
| Validation | 348 | 0 | 5,579 |
| Test | 612 | 31 | 10,131 |

# Model performance

- The corpus is difficult for models trained on other Russian GEC corpora.

| corpus | P | R | F0.5 | uncov., % |
|---|---|---|---|---|
| RULEC-GEC | 50.4 | 32.6 | 45.5 | 42.0 |
| RU-Lang8 | 60.8 | 37.9 | 54.2 | 48.8 |
| GERA | 74.3 | 47.0 | 66.6 | 33.7 |
| LORuGEC | 45.1 | 17.7 | 34.4 | 21.9 |

Table: Comparison of finetuned Qwen2.5-7B performance and difficult fraction (uncov., %) for different Russian GEC corpora. The model is tuned on the concatenation of Russian GEC data.

- It is not difficult to generate the corrections but hard to detect whether the correction is applicable.

# Fewshot motivation

- External finetuning performs even worse than zero-shot: LORuGEC deviates too much by error distribution.
- The validation set is too small for training.

| Setup | P | R | F |
|---|---|---|---|
| Zero-shot | 43.3 | 34.0 | 41.0 |
| ext. finetuning | 45.1 | 17.7 | 34.4 |
| ext.+LORuGEC finetuning | 50.1 | 37.9 | 47.1 |
| LORuGEC LORA finetuning | 48.6 | 42.6 | 47.3 |

Performance of Qwen2.5-7B on LORuGEC.

# Fewshot motivation

- External finetuning performs even worse than zero-shot: LORuGEC deviates too much by error distribution.
- The validation set is too small for training.
- What about in-context learning on it? No effect.

| Setup | P | R | F |
|---|---|---|---|
| Zero-shot | 43.3 | 34.0 | 41.0 |
| random, 1-shot | 44.4 | 28.6 | 40.0 |
| random, 5-shot | 47.2 | 30.2 | 42.4 |

Performance of Qwen2.5-7B on LORuGEC.

## Fewshot motivation

- External finetuning performs even worse than zero-shot: LORuGEC deviates too much by error distribution.
- The validation set is too small for training.
- What about in-context learning on it? No effect.

| Setup | P | R | F |
|---|---|---|---|
| Zero-shot | 43.3 | 34.0 | 41.0 |
| random, 1-shot | 44.4 | 28.6 | 40.0 |
| random, 5-shot | 47.2 | 30.2 | 42.4 |

Performance of Qwen2.5-7B on LORuGEC.

- But what in-context examples to select? Random samples are useless.
- Intuitively, examples for the same rule are more helpful.
- But how to detect them at inference time?

## In-context example selection

- We need a model that assigns similar vectors to sentences governed by the same rule.
- Common embedders do not have such property. They reflect meaning similarity, not syntactic or grammatical.
- We need an encoder trained on grammar-related task.

## In-context example selection: GECToR

- We need an encoder trained on grammar-related task.
- The natural solution is GECToR (Omelianchuk et al., 2020)

| source | target | label |
|---|---|---|
| BEGIN | BEGIN | APPEND_A |
| | A | |
| Boy | boy | LOWER |
| go | goes | VERB_3SG |
| | to | APPEND_to |
| school | school | KEEP |
| . | . | KEEP |

- It reduces grammar correction to sequence labeling.

# GECToR as sentence embedder

- GECToR provides meaningful representations for individual tokens in its last layer (before classification head).
- We need meaningful vectors for the whole sentence.
- Solution: **represent the sentence by embeddings of its most probable error positions**.
- We select up to 3 positions with error probability at least 0.1.
- If there are no such positions, use the most probable error position.
- **We reimplement GECTOR for Russian and English .**

# Further tuning GECToR

- Our goal is to assign similar vectors to sentences on the same rule.
- This might be achieved by contrastive tuning on the validation set using triplet loss:

$$L(h, h^+, h^-) = \max(\frac{\rho(h, h^+) - \rho(h, h^-) + \alpha}{t}, 0),$$

- Notation:
  - $h$ – current sentence vector,
  - $h^+$ – hard positive (closest example on the same rule),
  - $h^-$ – hard negative (closest example on another rule),
  - $\rho$ – cosine distance function,
  - $\alpha = 0.1$ – the margin.
- Indexes are created using FAISS and updated once an epoch.

# Results

**Main results**:

- Using GECToR for similar examples retrieval actually improves results.
- Further tuning of GECToR on rules annotation helps additionally.
- The improvement is consistent across models and their size.

| Setup | Qwen2.5-7B P | Qwen2.5-7B R | Qwen2.5-7B F0.5 | YandexGPT5-8B P | YandexGPT5-8B R | YandexGPT5-8B F0.5 | YandexGPT5-32B P | YandexGPT5-32B R | YandexGPT5-32B F0.5 |
|---|---|---|---|---|---|---|---|---|---|
| zero-shot | 43.3 | 34.0 | 41.0 | 66.4 | 51.0 | 62.6 | 76.5 | 66.7 | 74.3 |
| 1-shot, random | 44.4 | 28.6 | 40.0 | 67.8 | 48.6 | 62.8 | 78.3 | 71.0 | 76.7 |
| 5-shot, random | 47.2 | 30.2 | 42.4 | 68.5 | 56.3 | 65.6 | **83.9** | **79.2** | **83.0** |
| 1-shot, e5-base | 44.6 | 29.5 | 40.5 | 69.4 | 49.4 | 64.2 | 81.6 | 69.7 | 78.9 |
| 5-shot, e5-base | 47.0 | 31.8 | 42.9 | 68.8 | 56.8 | 66.0 | 81.8 | 72.2 | 79.7 |
| 1-shot, GECTOR | 50.2 | 35.8 | 46.5 | 69.9 | 53.9 | 66.0 | 81.9 | 72.8 | 79.9 |
| 5-shot, GECTOR | 54.3 | 41.7 | 51.2 | 70.0 | 62.4 | 68.3 | 82.7 | 76.7 | 81.4 |
| 1-shot, GECTOR+FT | *52.7* | *39.8* | *49.5* | *71.2* | *56.7* | *67.7* | *83.0* | *76.3* | *81.6* |
| 5-shot, GECTOR+FT | **59.3** | **46.2** | **56.1** | *73.1* | *65.5* | *71.4* | 83.5 | 78.1 | 82.3 |
| ext. finetuning | 45.1 | 17.7 | 34.4 | 67.0 | 35.4 | 56.9 | | NA | |
| ext.+LORuGEC ft. | *50.1* | 37.9 | 47.1 | **77.4** | **73.6** | **76.6** | | NA | |
| LORuGEC LORA ft. | 48.6 | *42.6* | *47.3* | 74.1 | 72.6 | 73.8 | | NA | |

# Results analysis

- Fewshot quality correlates with example retrieval quality.

| Retriever | acc. | top-5 recall | Qwen2.5-7B F0.5 1-shot | Qwen2.5-7B F0.5 5-shot | YandexGPT5-Pro F0.5 1-shot | YandexGPT5-Pro F0.5 5-shot |
|---|---|---|---|---|---|---|
| random | 2.3 | 10.3 | 40.0 | 42.4 | 76.7 | 83.0 |
| GECTOR | 31.7 | 49.3 | 46.5 | 51.2 | 79.9 | 81.4 |
| GECTOR+FT | 55.9 | 72.2 | 49.5 | 56.1 | 81.6 | 82.3 |

- Lexical errors are the hardest (results are for 5-shot GECTOR+FT):

| Category | Qwen2.5-7B P | Qwen2.5-7B R | Qwen2.5-7B F0.5 | YandexGPT5-Pro P | YandexGPT5-Pro R | YandexGPT5-Pro F0.5 |
|---|---|---|---|---|---|---|
| Grammar | 50.0 | 36.5 | 46.6 | 86.3 | 69.8 | 82.4 |
| Lexis | 46.7 | 22.6 | 38.5 | 85.0 | 54.8 | 76.6 |
| Punct. | 66.2 | 53.6 | 63.0 | 85.7 | 83.3 | 85.2 |
| Spelling | 55.2 | 44.9 | 52.8 | 80.9 | 77.4 | 80.2 |

# Results for other Russian corpora

- Results for RULEC-GEC (Rozovskaya et al.,2019) and ruLang-8(Trinh et al., 2021).

| | Qwen-2.5 7B Instruct | | | | | | YandexGPT-5 Lite 8B Instruct | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Setup | RULEC-GEC | | | RU-Lang8 | | | RULEC-GEC | | | RU-Lang8 | | |
| | P | R | F0.5 | P | R | F0.5 | P | R | F0.5 | P | R | F0.5 |
| zero-shot | 38.2 | 39.3 | 38.4 | 48.9 | 39.2 | 46.6 | 41.7 | 42.6 | 41.9 | 53.8 | 41.9 | 50.9 |
| random, 1-shot | 40.7 | *37.8* | 40.1 | 50.4 | 37.1 | 47.1 | 43.5 | 41.9 | 43.2 | 55.1 | 42.5 | 52.0 |
| random, 5-shot | 42.4 | 37.9 | 41.4 | 51.6 | 38.3 | 48.2 | 43.7 | 45.1 | 44.0 | 55.4 | 47.5 | 53.6 |
| gector, 1-shot | *41.8* | 37.6 | *40.9* | *53.7* | *38.8* | *49.8* | 45.0 | *42.5* | 44.5 | 56.9 | 43.5 | 53.6 |
| gector, 5-shot | 43.9 | 37.1 | 42.4 | *55.4* | 40.2 | 51.5 | 46.0 | 45.4 | 45.9 | *57.2* | *48.3* | *55.2* |
| gector+FT, 1-shot | 41.7 | 37.2 | 40.7 | 52.6 | 38.1 | 48.8 | *45.4* | 42.2 | *44.7* | *57.1* | *43.7* | *53.8* |
| gector+FT, 5-shot | *44.7* | **38.1** | *43.2* | 55.3 | **40.7** | *51.6* | *46.1* | **45.8** | *46.0* | 56.0 | 47.7 | 54.1 |
| finetuning | **52.2** | 31.2 | **46.0** | **61.7** | 37.2 | **54.5** | **57.3** | 38.9 | **52.4** | **66.3** | **48.5** | **61.8** |
| prev. SOTA | 70.5 | 29.1 | 54.8[2] | 73.7 | 27.3 | 55.0[1] | 70.5 | 29.1 | 54.8[2] | 73.7 | 27.3 | 55.0[1] |

- Untuned GECTOR is still helpful for fewshot (mostly improves precision).
- Tuning on LORuGEC doesn't provide further improvement (overfitting?).

# Results for English

- Results on BEA development set

| Method | few-shot method | k | Qwen2.5-7B | | | GPT4o-05-13 | | |
|---|---|---|---|---|---|---|---|---|
| Zero-shot | – | 0 | 36.2 | 43.4 | 37.5 | 34.2 | **52.6** | 36.8 |
| few-shot | random | 1 | 37.9 | 42.8 | 38.8 | 35.7 | 51.5 | 38.0 |
| few-shot | random | 5 | 38.4 | 43.6 | 39.4 | 37.2 | 49.0 | 39.1 |
| few-shot | GECTOR | 1 | 39.1 | 44.4 | 40.1 | 37.2 | 52.0 | 39.4 |
| few-shot | GECTOR | 5 | 40.0 | 46.0 | 41.1 | **39.4** | 51.5 | **41.4** |
| LLM finetuning | – | 0 | **53.4** | **48.8** | **52.4** | NA | NA | NA |

- GECToR (retrained on cLang-8) still improves in-context learning.
- Fewshot (and LLM) performance in general is poor.

# Conclusions

- Contributions:
  - We create a new rule-oriented corpus for Russian GEC.
  - We demonstrate the usefulness of GECToR-like models for few-shot example selection .
  - We suggest a method for further GECTOR tuning on error-type annotated corpus.
- Limitations:
  - Limited examples of successful GECTOR tuning (didn't work for BEA).
  - GECTOR-like models exist only for a few languages.
- Future extensions:
  - Extend to other languages and corpora.
  - Extend to other task where task-induced similarity differs from surface similarity (Math, coding, ...).

# (Recent) related work

- Peng et al. *Encode Errors: Representational Retrieval of In-Context Demonstrations for Multilingual Grammatical Error Correction* (ACL Findings, 2025).
  - Transforms internal LLM states via PCA to extract grammar-related representations.
  - Significant improvement over random few-shot for English, German and Estonian.
- Liu et al. *Unraveling the Mechanics of Learning-Based Demonstration Selection for In-Context Learning* (ACL 2025).
- Chen et al. *Retrieval-style In-Context Learning for Few-shot Hierarchical Text Classification* (TACL 2025).
- Peng et al. *Enhancing Input-Label Mapping in In-Context Learning with Contrastive Decoding* (ACL 2025).

# Links

- Data: `https://github.com/ReginaNasyrova/LORuGEC`
- Code: `https://github.com/AlexeySorokin/LORuGEC`
- GECToR code (for Russian):
  `https://github.com/ReginaNasyrova/RussianGEC_SeqTagger`
- GECTOR paper (for Russian): R. Nasyrova, A. Sorokin *Grammatical Error Correction via Sequence Tagging for Russian* (accepted to ACL-SRW 2025).

## Acknowledgements


- Eva Gogua, Kristina Mustakova and Elina Ozhogova – for data annotation.
- Marina Kosheleva – for help with YandexGPT.


# Thank you for your attention!

# Danke schön! Спасибо!