Adapting LLMs for Minimal-edit Grammatical Error Correction

Adam Mickiewicz University Poznań

CENTER FOR ARTIFICIAL INTELLIGENCE

Ryszard Staruch, Filip Graliński, Daniel Dzienisiewicz
Adam Mickiewicz University, Center for Artificial Intelligence

Introduction

Minimal-edit vs fluency-edit grammatical error correction (GEC)

Source text	I can recommed you to my uncle's company to do a job.
Fluency-edit output	I can recommend you for a job at my uncle's company.
Minimal-edit output	I can recommend you to my uncle's company to get a job.

The problem

LLMs tend to produce fluency-edit outputs.

The solution

Adapting LLMs for minimal-edit GEC through:

- Dataset detokenization
- Data augmentation
- Training schedule method

Dataset detokenization

Tokenized vs detokenized GEC examples

Tokenized sentences	Detokenized sentences
In the future , I 'll become a journalist.	In the future, I'll become a journalist.
She 's 9 or 10 years old . I do n't really know .	She's 9 or 10 years old. I don't really know.
I am a reliable , easy - going person .	I am a reliable, easy-going person.

The problem

Datasets are available only in tokenized format, requiring models to learn tokenization patterns.

The solution

Detokenization of the following datasets:

- W&I+LOCNESS
- FCE
- CoNLL-2014
- JFLEG

Detokenization approaches

- Rule-based sentence splitting
- Sacremoses detokenizer + manual fixes
- Llama 3.1-70B-3.1

Incorrect annotations in datasets

The Llama model occasionally (in fewer than 10% of examples) corrected errors in human annotations.

Examples:

- ...to a cafe and ~~and~~ I drank a drink.
- I recommend you to practise any sport...
- She is one of the ones that...

Experiments with detokenized datasets

Model	Dataset type	W&I+LOCNESS dev		
		Precision	Recall	$F_{0.5}$
Qwen-2.5-1.5B	tokenized	59.00	38.48	53.31
	detokenized	57.90	42.10	53.86
Llama-3.2-3B	tokenized	63.31	47.29	59.29
	detokenized	63.34	47.52	59.39
Gemma-2-9b	tokenized	68.84	55.90	65.79
	detokenized	68.84	56.40	65.93

Findings

- Models can learn the tokenization schema alongside the GEC task.
- The detokenization process revealed annotation errors in high-quality datasets.

Data augmentation

Typical GEC training for neural models trained from scratch involves removing unedited examples to improve recall. For LLMs, which already generalize well, we do the opposite.

Our approach

Augment the dataset by duplicating corrected sentences as both input and target.

Original pair
Source
Alice have a cat.
Target
Alice has a cat.

Augmented pair
Source
Alice has a cat.
Target
Alice has a cat.

Experiments (Gemma 2 9b)

Dataset processing approach	W&I+LOCNESS dev		
	Precision	Recall	$F_{0.5}$
Only erroneous pairs	60.74	58.79	60.34
Original pairs	68.99	57.12	66.24
Original pairs + augmented pairs	71.42	53.42	**66.92**

Findings

- Removing unedited pairs worsens results compared to using the original datasets.
- Providing augmented pairs enables control over the precision-recall trade-off.

Training schedule method

To control the precision-recall trade-off during training, we introduce a novel training schedule method.

How it works?

Train the model only on erroneous examples. → Then, fine-tune it on correct examples using a lower learning rate.

The model first learns to correct grammar. → Then, it learns when not to correct, achieving better precision.

Our best model training schedule

Dataset: FCE with data augmentation
Epochs: 1
Learning rate: 5e-6

→

Dataset: W&I+LOCNESS (only erroneous examples)
Epochs: 1
Learning rate: 5e-6

→

Dataset: W&I+LOCNESS (only correct examples)
Epochs: 1
Learning rate: 3e-7

Experiments (Gemma 2 9b)

Lowered learning rate for the unedited examples	W&I+LOCNESS dev		
	Precision	Recall	$F_{0.5}$
1e-7	65.10	58.33	63.62
2e-7	69.22	56.40	66.21
3e-7	73.52	50.10	**67.23**
4e-7	76.74	43.49	66.57
5e-7	78.88	31.78	60.85

Findings

- Choosing the proper learning rate enables control over the precision–recall trade-off.
- The parameter is very sensitive — it significantly affects the model's performance.

Results on the test sets

Model	CoNLL-2014 test			W&I+LOCNESS test		
	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$
Llama-2-13b (Omelianchuk, et al., 2024)	77.30	45.60	67.90	74.60	67.80	73.10
Mistral-7b-EPO (Liang et al., 2025)	76.71	52.56	**70.26**	78.16	68.07	75.91
Gemma-2-9b-Augmentation	73.80	**56.16**	69.43	74.86	**71.35**	74.13
Gemma-2-9b-Training-Schedule	75.74	51.47	69.24	79.87	68.90	77.41
Llama-2-13b-Training-Schedule	71.07	50.11	65.59	74.10	67.54	72.69
Gemma-2-27b-Training-Schedule	**77.38**	47.88	68.89	**82.28**	67.03	**78.70**

Findings

- Our methods enable training minimal-edit GEC models with a focus on precision.
- Choosing a modern LLM leads to better final results.
- Our models achieve new SOTA single-model results on the W&I+LOCNESS test set.

Conclusions

- The detokenization process did not improve model performance.
- There are annotation errors even in high-quality datasets.
- Both of our methods can be leveraged to adapt LLMs for minimal-edit GEC.

Detokenized datasets

Code