MultiGEC-2025

A shared task in Multilingual Grammatical Error Correction

A. Masciolini, A. Caines, O. De Clercq, J. Kruijsbergen, M. Kurfalı, R. Muñoz Sánchez, E. Volodina, R. Östling and many, many others

March 5, 2025

The shared task in short

MultiGEC-2025

- Grammatical Error Correction (GEC)
- multilingual (12 European languages)
- text-level
- two tracks:
  1. “minimal edits”
  2. “fluency edits”

What is GEC?

Grammatical Error Correction is *sequence-to-sequence task* where:

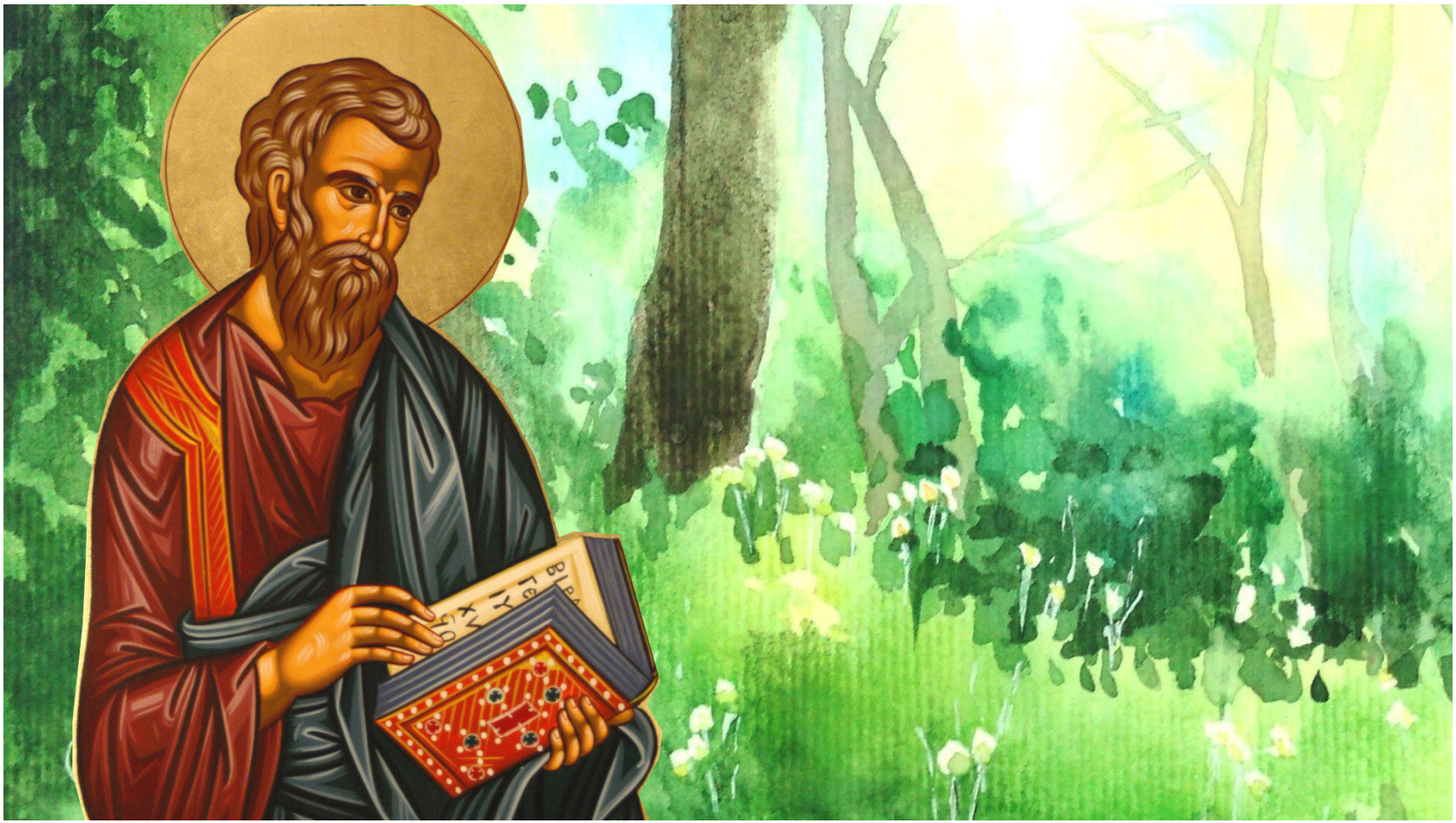
- **input**: a text, potentially ungrammatical, typically written by a learner
- **output**: a normalized version of the same text, aka *correction hypothesis*, which can be
  - *minimal* or
  - *fluency-edited*

Example

original	**normalized (minimal)**	**normalized (fluency)**
My moter became very sad, no food.	My *mother* became very sad, *and ate* no food.	My *mother was* very *sad and refused to eat*.
Min mama bliv väldigt ledsen, ingen mat.	Min *mamma blev* väldigt ledsen, *och åt ingen mat*.	Min *mamma blev* väldigt ledsen *och slutade äta*.
Mia mama era tanto triste, mangiava niente.	Mia *mamma* era tanto triste *e non* mangiava niente.	Mia *madre* era tanto triste *che aveva smesso di mangiare*.
Mi mama era tan triste, no comia.	Mi *mamá estaba muy* triste, *no comía*.	Mi *mamá estaba muy* triste *y no comía nada*.

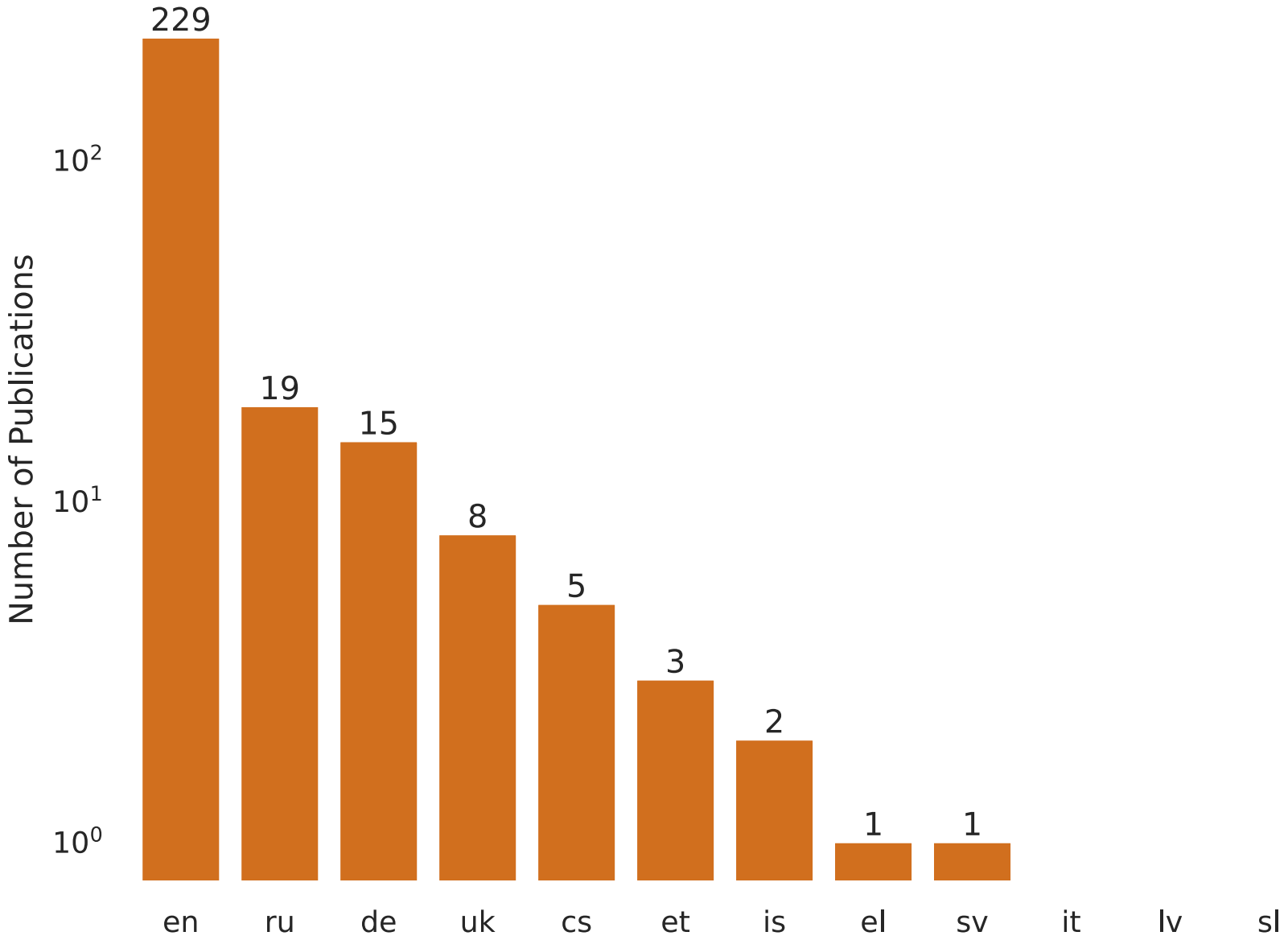
(in the shared task, this is done at the level of full texts)

Why multilingual?

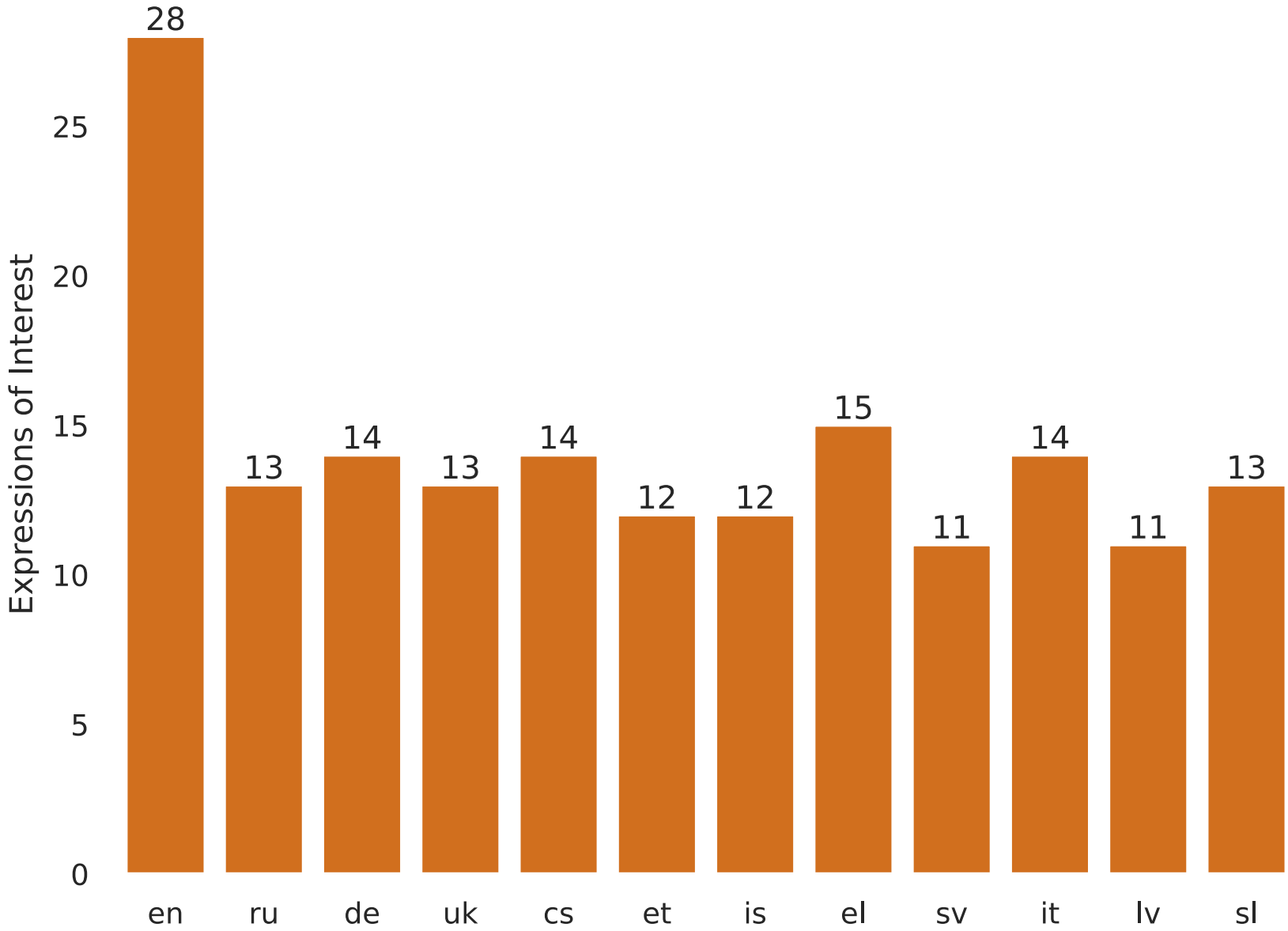


Why multilingual?

Why multilingual?



Why multilingual?



The MultiGEC dataset

lang	subcorpus	n. essays	learners	minimal	fluency	peculiarities
cs	NatWebInf	6167	**L1** (web)	✓		
cs	Romani	3599	**L1** (Romani background)	✓		
cs	SecLearn	2407	**L2**	✓		
cs	NatForm	391	**L1** (students)	✓		
en	Write & Improve	5050	**L2**	✓		separate download
et	EIC	258	**L2**	✓	✓	
et	EKIL2	1503	**L2**		✓	
de	Merlin	1033	**L2**	✓		
el	GLCII	1289	**L2**	✓		
is	IceEC	176	**L1**		✓	pre-tokenized
is	IceL2EC	193	**L2**		✓	pre-tokenized; text fragments
it	Merlin	813	**L2**	✓		
lv	LaVA	1015	**L2**	✓		
ru	RULEC-GEC	6043	**mixed** (L2 + heritage)	✓	✓	pre-tokenized; text fragments; separate download
sl	Solar-Eval	109	**L1** (students)	✓		
sv	SweLL_gold	502	**L2**	✓		
uk	UA-GEC	1872	**mixed** (crowdsourced)	✓	✓	

Baseline

Automatic evaluation

metric	**characteristics**
ERRANT-based $F_{0.5}$ score	reference-based; winning metric for **track 1**
GLEU score	rewards fluency, but still reference-based
Scribendi score	reference-free (LM[1]-based); winning metric for **track 2**

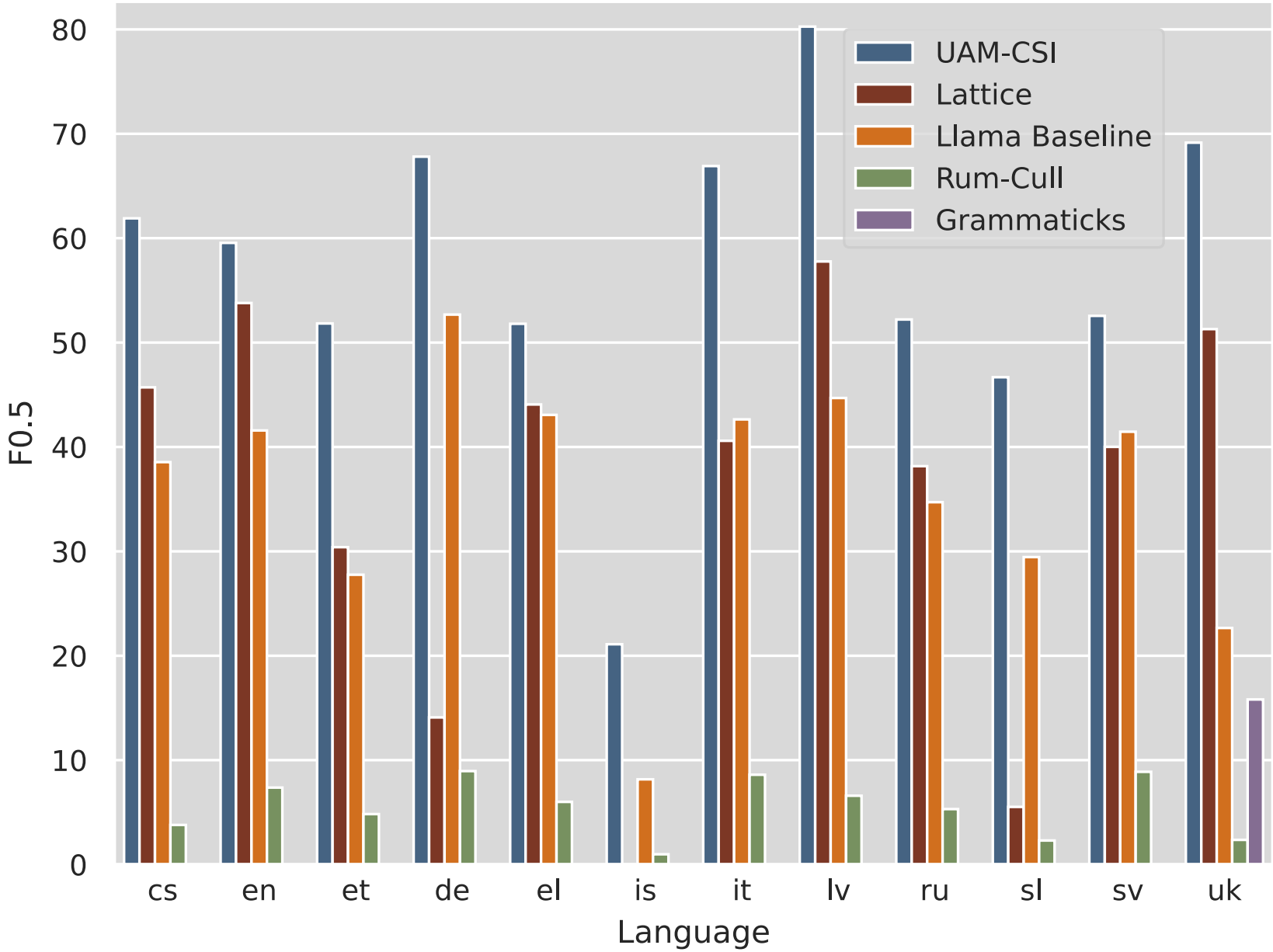
...all adapted to work on **full texts** in all **12 languages**!

[1] Gemma 2

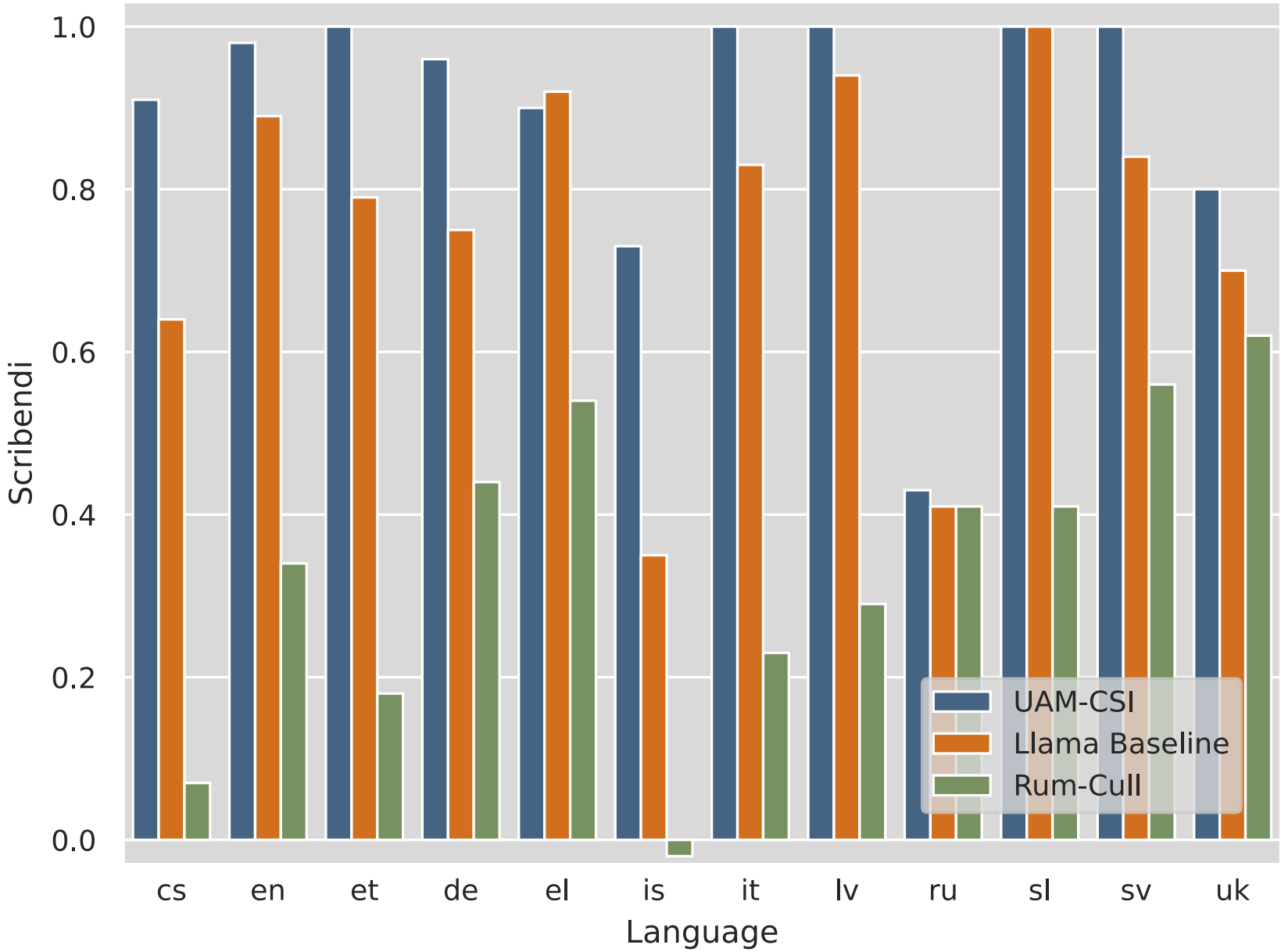
System submissions

team	languages	track 1	track 2	system
UAM-CSI	all 12	✓	✓	fine-tuned Gemma 2
Lattice	11	✓		fine-tuned LlaMA 3
Lattice	Slovene	✓		XLM-RoBERTa pipeline
Rum-Cull	all 12	✓	✓	?
Grammaticks	Ukrainian	✓		?

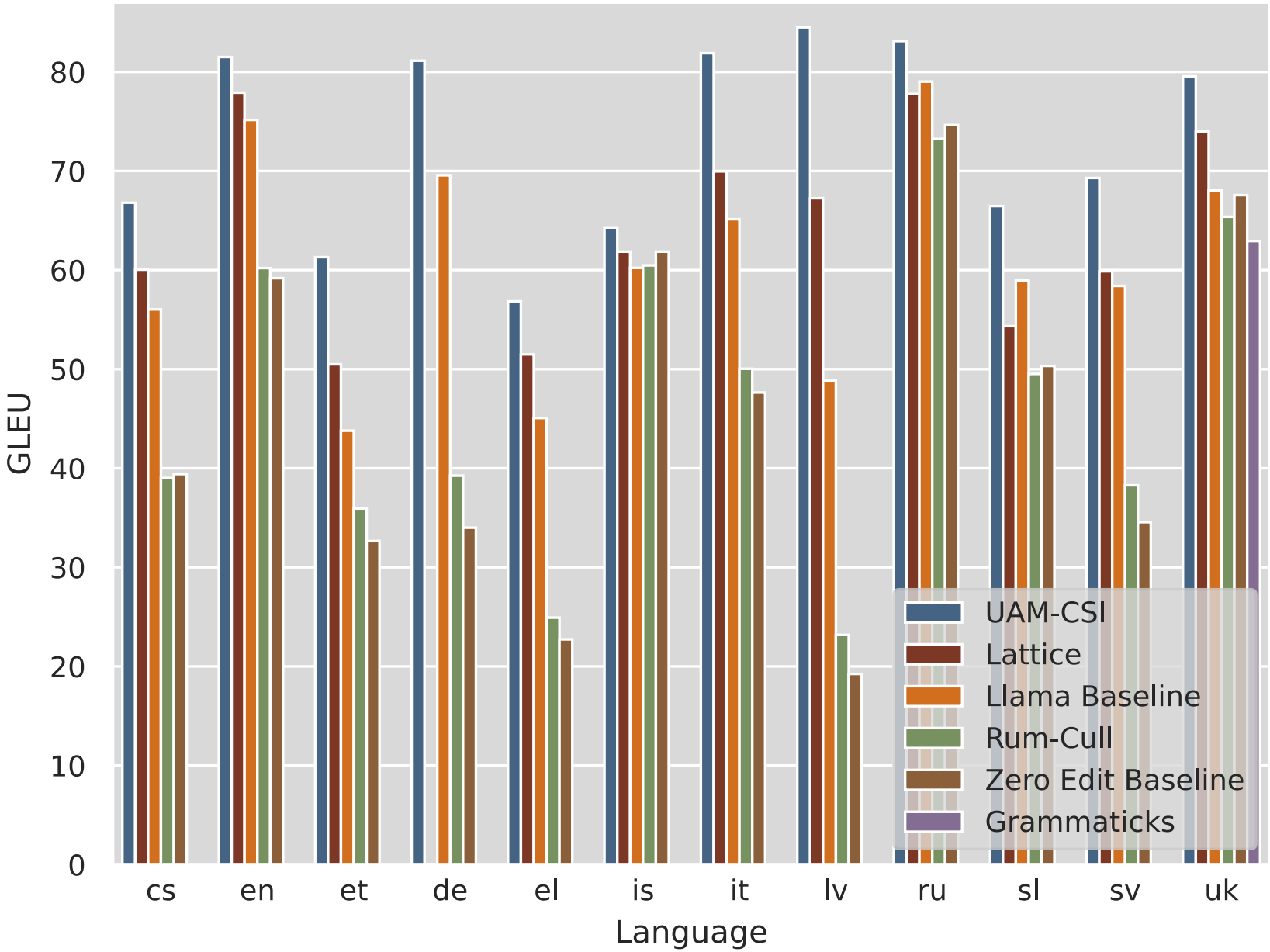
Results (track 1)



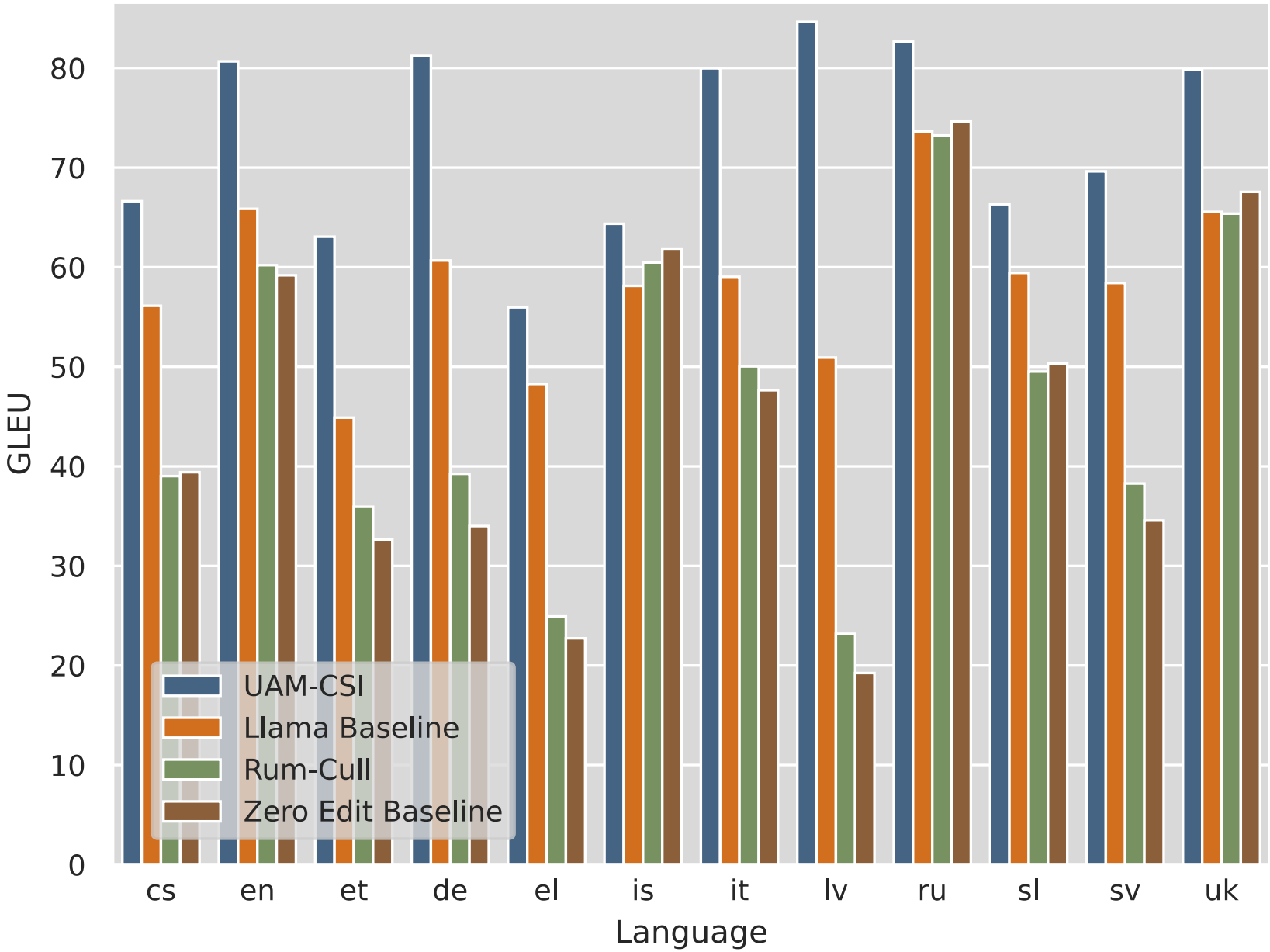
Results (track 2)



GLEU scores (track 1)



GLEU scores (track 2)



Manual evaluation (preliminary)

- 5 languages (English, German, Italian, Russian and Swedish)
- manual inspection of all outputs for one challenging essay per language
- confirms the ranking
- further work needed to know the extent to which scores are cross-lingually comparable

Key takeaways

- the overall winner is team UAM-CSI
- in track 1, team Lattice ranks second for most languages
- the baseline was harder to beat than expected
- some languages (Icelandic & Russian) proved extra challenging for all models

Reflections

4 submissions vs. (∼**40 + 10**) requests for access to the dataset.
How to favor broader participation?

Reflections

4 submissions vs. (~**40** + **10**) requests for access to the dataset.
How to favor broader participation?

- **timeline**: longer development phase, shorter test phase

Reflections

4 submissions vs. (~**40 + 10**) requests for access to the dataset.
How to favor broader participation?

- **timeline**: longer development phase, shorter test phase
- **baseline**: LLM baselines may discourage submissions of supervised system

Reflections

4 submissions vs. (~**40 + 10**) requests for access to the dataset.
How to favor broader participation?

- **timeline**: longer development phase, shorter test phase
- **baseline**: LLM baselines may discourage submissions of supervised system
- **dataset**:
  - number of languages/subcorpora vs. ease of access
  - test sets: to release or not to release?

Reflections

4 submissions vs. (~**40 + 10**) requests for access to the dataset.
How to favor broader participation?

- **timeline**: longer development phase, shorter test phase
- **baseline**: LLM baselines may discourage submissions of supervised system
- **dataset**:
  - number of languages/subcorpora vs. ease of access
  - test sets: to release or not to release?
- **evaluation**: advanced metrics vs. practical constraints

The future

- **dataset**:
  - v1.1 enhancing cross-subcorpus consistency
  - extension of MultiGEC with additional references and/or new languages

The future

- **dataset**:
  - v1.1 enhancing cross-subcorpus consistency
  - extension of MultiGEC with additional references and/or new languages
- **evaluation**:
  - larger-scale manual assessment of shared task submissions
  - further work on cross-lingually applicable metrics
  - further automation

The future

- **dataset**:
  - v1.1 enhancing cross-subcorpus consistency
  - extension of MultiGEC with additional references and/or new languages
- **evaluation**:
  - larger-scale manual assessment of shared task submissions
  - further work on cross-lingually applicable metrics
  - further automation
- **new MultiGEC systems** during the task's open phase

To learn more about...

- **the state of GEC in the 12 MultiGEC languages prior to the shared task**: A. Masciolini, A. Caines, O. De Clercq, J. Kruijsbergen, M. Kurfalı, R. Muñoz Sánchez, E. Volodina, R. Östling, K. Allkivi, Š. Arhar Holdt, I. Auzina, R. Dargis, E. Drakonaki, J. Frey, I. Glišić, P. Kikilintza, L. Nicolas, M. Romanyshyn, A. Rosen, A. Rozovskaya, K. Suluste, O. Syvokon, A. Tantos, D. Touriki, K. Tsiotskas, E. Tsourilla, V. Varsamopoulos, K. Wisniewski, A. Žagar, and T. Zesch. *An overview of Grammatical Error Correction for the twelve MultiGEC-2025 languages*. Gothenburg, Sweden, 2025
- **the shared task**: A. Masciolini, A. Caines, O. De Clercq, J. Kruijsbergen, M. Kurfalı, R. Muñoz Sánchez, E. Volodina, and R. Östling. *The MultiGEC-2025 shared task on Multilingual Grammatical Error Correction at NLP4CALL*. In R. Muñoz Sánchez, David Alfter, Jelena Kallas, and E. Volodina, editors, Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning, Tallin, Estonia, 2025

To learn more about...

- **the MultiGEC dataset**:
  - **paper**: A. Masciolini, A. Caines, O. De Clercq, J. Kruijsbergen, M. Kurfalı, R. Muñoz Sánchez, E. Volodina, R. Östling, K. Allkivi, Š. Arhar Holdt, I. Auzina, R. Dargis, E. Drakonaki, J. Frey, I. Glišić, P. Kikilintza, L. Nicolas, M. Romanyshyn, A. Rosen, A. Rozovskaya, K. Suluste, O. Syvokon, A. Tantos, D. Touriki, K. Tsiotskas, E. Tsourilla, V. Varsamopoulos, K. Wisniewski, A. Žagar, and T. Zesch. *Towards better language representation in Natural Language Processing – a multilingual dataset for text-level Grammatical Error Correction*. To appear in the International Journal of Learner Corpus Research, 2025
  - **dedicated website**: spraakbanken.github.io/multigec-2025
  - **resource page**: doi.org/10.23695/h9f5-8143
  - **download**: lt3.ugent.be/resources/multigec-dataset