

# Exploring limitations and risks of LLM-based GEC for Indigenous languages

Who? Flammie A Pirinen 🇫🇮 , Linda Wiechetek 🇺🇸

From? 🇫🇮 UiT, Norgga árktaš universitehta, Divvun

When? February 28, 2025 at Computel 8

# Introduction & Backgrounds

- 1 We have been developing rule-based GEC for Indigenous languages for the last decade-ish
- 2 LLMs are currently being sold as silver bullet for all NLP tasks
- 3 We set out to test where they are now at for GEC in Indigenous language context
- 4 Trying to find out good and bad use cases and ethical use

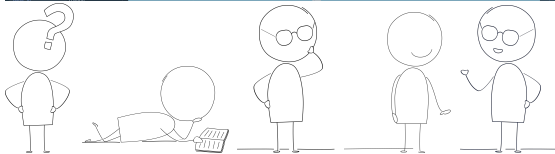
# Who we are

## Divvun at UiT Norway's Arctic university

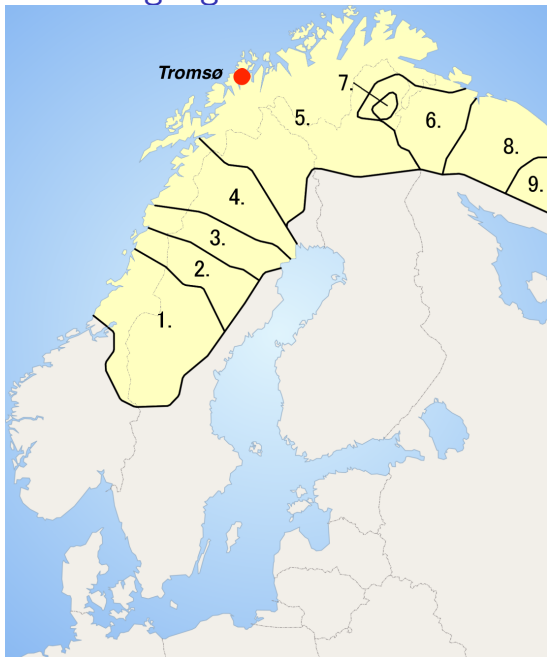
Forbindelse: Tromsø (407754)  
Versjon #29  
Update population in Norway. Large changed bounding box due to Norwegian sub-Antarctic dependent territories.  
Rudolf 9 måneder siden av NKA  
Changelog #1028285

**Egenskaper**

address_level	7
alt_name_en	Tromsø
boundary	administrative
name	Tromsø
name_ar	ترومسو
name_b	Tromsø
name_fa	ترومسو
name_ja	トロムソ
name_ko	트로مس오
name_lv	Tromsøda
name_n	Tromsø
name_nb	Tromsø
name_no	Tromsø
name_pt	Tromsø
name_ru	Тромсø
name_sv	Tromsø
name_zh	特隆斯奧



# Sámi languages



# Features of Sámi languages especially with regard to GEC

- Uralic languages, not related to Indo-European
- Morphological complexity: more than average Indo-European, less than poly-synthetic
- Young and developing orthographies and standards, active literary culture (media, administration, fiction, translations)
- Language is taught in schools: many of our users are students often dependent on functional language tools for homework and exams

# What is GEC?

- Grammatical error correction concerns correcting text based on *grammar*
- Mainly concerns on fixing correctly spelled real words in wrong context, i.e. not misspellings that result in a non-word
- e.g. subject does not agree with verb form, adposition's complement is in wrong case

# Importance and quality of GEC

- GEC is part of building of a language standard / revitalisation support for writers
- There are both L1 and L2 writers, and GEC is commonly used
- so: GEC must have **high precision** and follow the language standard that is supported by authorities / community
- Hypothesis: LLM-based GEC struggles by preferring what is common in data over what is the language standard
- Also: having LLM as an responsible authority of language norms?

# Experiment

We randomly selected 100 sentences from our error corpus and gave them to ChatGPT and our rule-based grammar checker GramDivvun. Sentences contained between zero and many errors.



# ChatGPT



The sentence you provided has a few areas that need correction for proper grammar in North Sámi:

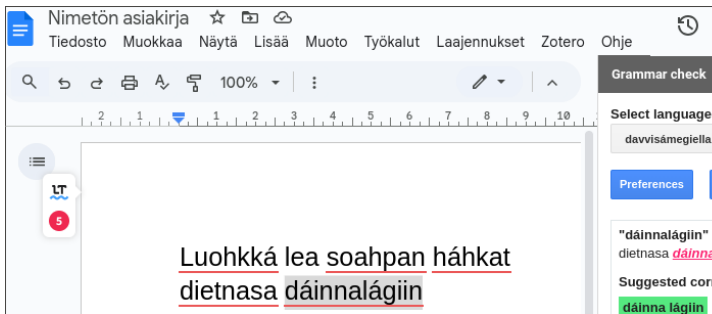
1. **Gii livčče** should be **Gii livččii** (Who was)
2. **jáhkkán** should be **jáhkán** (waiting)
3. **Guorosnaga** should be **Guorusnaga** (Suddenly)
4. **manai** should be **manai** (went, no change needed)
5. **ja dál diekkár** should be **ja dál diehtá** (and now knows)

Here is the corrected version:

"- Gii livččii jáhkán. Guorusnaga manai ja dál diehtá."

This translates to: "Who was waiting. Suddenly went and now knows."

# GramDivvun



The screenshot shows a Google Docs interface with a document titled "Nimetön asiakirja". The text in the document is "Luohkká lea soahpan háhkat dietnasa dáinnalágiin". Red lines underline the words "soahpan", "háhkat", "dietnasa", and "dáinnalágiin". On the right side, the "Grammar check" sidebar is visible, showing "Select language" set to "davvisámegiella" and "Suggested corrections" for "dáinna lágiin".

Figure: The red lines are GoogleDocs default English grammar checking 🙄

# Results

System	Precision	Recall	$F_{0.5}$
GramDivvun	58 %	60 %	0.58
ChatGPT	17 %	13 %	0.16

**Table:** Precision, recall and  $F_{0.5}$  scores of the systems we tested. NB: One should **not** use  $F_1$  with GEC, since the emphasis is on precision.

# Conclusions

- LLMs are not yet at the level of rule-based GEC at all
- in general—as in, for English—they improve at rapid pace for now
- eventually will get rolled out in more commercial products
- → we need to ensure we aren't getting forced to pay for low quality AI GEC
- For language revitalisation, building of a language standard, etc., LLM cannot take authority or responsibility
- Failures in North Sámi LLM error correction are critical and the output cannot be trusted (which is a contradiction to its use ...)

# Giitu / Thank you 🙏, Gažaldagat / Questions ?

[feedback@divvun.no](mailto:feedback@divvun.no)  
<https://divvun.no>



Pirinen, F., Moshagen, S., and Hiovain-Asikainen, K. (2023).

GiellaLT — a stable infrastructure for Nordic minority languages and beyond.

In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.



Wiechetek, L., Moshagen, S., Gaup, B., and Omma, T. (2019).

Many shades of grammar checking—launching a constraint grammar tool for north sámi.