language and translation technology team

Getting More from Small LLMs and Limited Data

Fuzzy-Match Retrieval and Candidate Scoring for Multitask Sorbian NLP

Thomas Moerman, Arda Tezcan

LT3, Language and Translation Technology Team, Ghent University

We serve five Sorbian tasks with one small LLM (2B) and were joint winner at WMT26.
It translates with retrieved examples and picks answers instead of writing them.

1 The task

Can one small model handle five tasks in Upper and Lower Sorbian?

Translation	writes	Question answering	picks	Spell checking	picks	Grammar checking	picks	Maths reasoning	writes
de *So erleben die Kinder…* hsb *Tak dožiwjeja dźěći…* six directions with German		*Džakuju so!* → ***Tu prošu!*** "Thank you!" "Here you are!"		*…wo twarški serwis…* → ***twarski*** flag the word, give its spelling		*…wot 1. awgust 2022…* → ***awgusta*** the date needs the genitive		*Kelko kulow trjebaš cyłkownje?* → ***3*** a word problem, a number back	
69.57 chrF++	2nd	**65.05** accuracy	1st	**76.11** exact match	2nd	**78.32** exact match	2nd	**30.40** accuracy	1st

Official scores, hidden test. QA is 7.7 points ahead of the next system, maths 1.4. LT3 and HeyBusan are joint winners (22 rank points each).

2 Translation

Only training with authentic fuzzy matches keeps improving.

A fuzzy match: a similar sentence with its translation.

Translate from German to Upper Sorbian

retrieved match, an authentic pair

German Die Kinder und die Eltern haben das Sprachangebot gut angenommen.

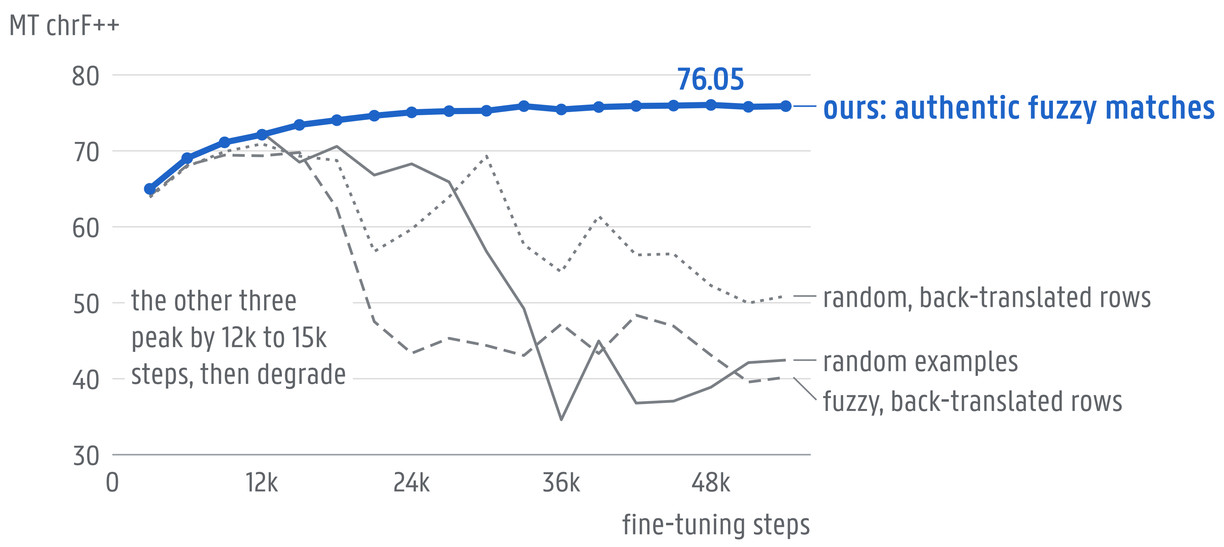
Upper Sorbian Dźěći a starši su rěčny poskitk derje přiwzali.

sentence to translate

German **So erleben die Kinder Freude mit und an der Sprache.**

Upper Sorbian **…**

Training 0 or 1 match per row, from authentic pairs only.
Inference 3 matches, from authentic and back-translated pairs.



Development sets, four fine-tuning recipes, identical inference (three fuzzy matches from authentic pairs).

- Back-translated pairs cost 6.3 chrF++ as training rows. In the retrieval pool they do no harm and add a little: 66% of the matches for German sources are back-translated.
- Examples do not help the untuned model (25.8 chrF++). Fine-tuned with fuzzy matches, it reaches 76.1.

3 The other tasks

Letting the model pick among candidates lifts QA by 15 points and spelling by 10.

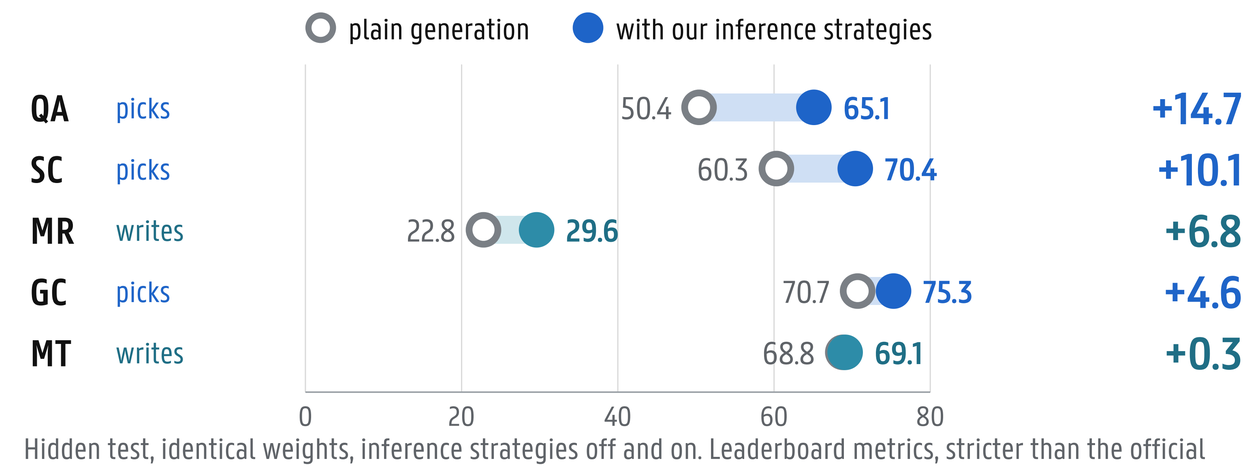
How picking works, on a spelling example:

1. The model flags a word: *…wo twarški serwis…*
2. A dictionary proposes corrections: *twarski*
3. The model scores each full sentence in one pass:

…wo twarški serwis…	**−4.39** unchanged
…wo twarski serwis…	**−3.84 picked**

4. The higher score wins, here the gold answer.

Answer options and grammar forms are picked the same way.



Hidden test, identical weights, inference strategies off and on. Leaderboard metrics, stricter than the official ones for SC and GC.

- No training data existed for these four tasks. We synthesised about 330,000 items from dictionaries, monolingual text and translated maths problems.
- A 2B model can judge whether a form is right, but often cannot write the right form itself.

1 Teach retrieval with clean data.
Train on authentic fuzzy matches. Keep back-translations for the retrieval pool.

2 Compare recipes at convergence.
Fuzzy and random examples tie early and end 3.6 chrF++ apart.

3 Let small models pick, not write.
Scoring listed candidates gave our largest gains.

Limits. One track, two related languages, one 2B model. Automated metrics only. Some inference thresholds were fitted on the development sets.

GHENT UNIVERSITY

Code, data, models

github.com/TomMoeras/wmt26-lt3-sorbian

hf.co/collections/TomMoeras/wmt2026-multitasksorbiannlp-lt3

thomas.moerman@ugent.be arda.tezcan@ugent.be

Sorbian data: WITAJ-Sprachzentrum. Compute: VSC, funded by Ghent University, FWO and the Flemish Government.

TransTech

EMNLP2026 BUDAPEST