Findings of the WMT 2025 Shared Task LLMs with Limited Resources for Slavic Languages: MT and QA

Shu Okabe[1,2] **Daryna Dementieva**[1,2] **Marion Di Marco**[1,2] **Lukas Edman**[1,2]
Kathy Hämmerl[1,2] **Marko Měškank**[3] **Anita Hendrichowa**[3] **Alexander Fraser**[1,2,4]

[1]Technische Universität München [2]Munich Center for Machine Learning
[3]WITAJ-Sprachzentrum [4]Munich Data Science Institute

Correspondence: shu.okabe@tum.de, daryna.dementieva@tum.de

Abstract

We present the findings of the WMT 2025 Shared Task *LLMs with Limited Resources for Slavic Languages*. This shared task focuses on training LLMs using limited data and compute resources for three Slavic languages: Upper Sorbian (hsb), Lower Sorbian (dsb), and Ukrainian (uk), with the objective to develop and improve LLMs for these languages. We consider two tasks which are to be evaluated jointly: Machine Translation (MT) and Multiple-Choice Question Answering (QA).

In total, three teams participated in this shared task, with submissions from all three teams for the Sorbian languages and one submission for Ukrainian. All submissions led to an improvement compared to the baseline Qwen2.5-3B model through varying fine-tuning strategies. We note, however, that training purely on MT degrades original QA capabilities. We also report further analyses on the submissions, including MT evaluation using advanced neural metrics for Ukrainian, as well as manual annotation and comparison to the current Sorbian machine translator.

1 Introduction

For a large majority of the world's languages, only limited resources are available for training NLP tools, but modern large language models (LLMs) need large amounts of both labelled and unlabelled data to function well. Improving the coverage of low-resource languages in LLMs is an active research area. Recent examples include Nag et al. (2025) for Indic languages and Tonja et al. (2024) for Ethiopian languages. Similarly, there is active work on low-resource machine translation, with recent shared tasks and datasets covering translation for low-resource Indic languages (Pakray et al., 2024), Creole languages (Robinson et al., 2024) as well as Indigenous Languages of the Americas (De Gibert et al., 2025).

Although commercial LLMs increasingly show high performance on both general tasks and machine translation, specialised MT models are still typically required for best results. Our challenge to participants in this shared task is to build a model under low-resource conditions to jointly optimise machine translation and question answering. We aim to study potential synergy effects between these two tasks, as well as to explore whether optimising for one task in a low-resource setting will negatively impact the other task. We are one of the first WMT shared tasks to focus on joint optimisation of Machine Translation (MT) and Question Answering (QA). The Multilingual Instruction Shared Task in the same year took a similar approach, but allowed significantly larger models.

Our task focuses on three Slavic languages: Upper Sorbian, Lower Sorbian, and Ukrainian. Thus, we aim to highlight both truly low-resource settings and mid-resource language scenarios in the context of modern language technologies. As our goal is to evaluate LLMs as general-purpose tools for a given language, we designed our setup to mirror widely adopted benchmarks such as GLUE (Wang et al., 2019) and MMLU (Hendrycks et al., 2021). Since Sorbian languages and Ukrainian currently lack such comprehensive language understanding benchmarks, we approximated a multitask evaluation by selecting two representative tasks: Machine Translation and Question Answering.

Previous iterations of the WMT Shared Tasks on translating low-resource languages (Weller-Di Marco and Fraser, 2022; Libovický and Fraser, 2021; Fraser, 2020) compared supervised and unsupervised translation in various data settings. For both Sorbian languages, the WITAJ-Sprachzentrum provided new machine translation and question answering datasets for this shared task. For Ukrainian, the MT portion of the dataset corresponds to that of the WMT 2025 general translation

Proceedings of the Tenth Conference on Machine Translation, pages 503–519
November 8-9, 2025

task,[1] and the QA portion is based on the dataset from the UNLP 2024 Shared Task on fine-tuning LLMs for Ukrainian (Romanyshyn et al., 2024).

Our research questions are as follows:

- In a low-resource scenario, how does training a model for machine translation impact its performance on a secondary task such as question answering?
- Is it possible to improve capabilities on both machine translation and question-answering in a small LLM?

This article is structured as follows: Section 2 describes the shared task rules, while Section 3 details the MT and QA datasets provided for both development and test phases. Section 4 presents the systems devised by the three participating teams. Section 5 displays the official leaderboard for the primary submissions in all three tracks. Section 6 analyses the model outputs with some additional experiments.

2 Shared Task Description

2.1 Languages

Upper Sorbian (ISO code: `hsb`; Glottocode:[2] `uppe1395`) and Lower Sorbian (`dsb`; `lowe1385`) are minority languages spoken in the eastern part of Germany in the federal states of Saxony and Brandenburg, with only 30k and 7k native speakers, respectively. As western Slavic languages, Upper and Lower Sorbian are closely related to Polish and Czech. There is an active language community working on the preservation of these languages, namely the WITAJ-Sprachzentrum[3] (WITAJ Language Center) who also provided parts of the data used in this shared task. Previously, the WMT Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Weller-Di Marco and Fraser, 2022; Libovický and Fraser, 2021; Fraser, 2020) focused on both languages.

Ukrainian (`ukr`; `ukra1253`), spoken by approximately 40 million L1 speakers worldwide, is considered a mid-resource language in NLP with already several language-specific pre-training corpora (Chaplynskyi, 2023) and LLMs (Yukhymenko et al., 2025) available. With a significant number of Ukrainians currently living abroad, the demand for high-quality machine translation systems to support integration into new environments is greater than ever. Machine translation, complemented by cross-lingual knowledge transfer and robust question answering capabilities, can play a crucial role in addressing this need.

2.2 Task Description

Our main goal is to observe the synergy between the two different tasks in an LLM. Therefore, models are tested *jointly* on two tasks: **machine translation** and **multiple-choice question answering**. All participating systems must submit outputs for both tasks of a track.

For Machine Translation, we focus on the following translation directions:

- German to Upper Sorbian (de→hsb)
- German to Lower Sorbian (de→dsb)
- English to Ukrainian (en→uk)
- Czech to Ukrainian (cs→uk)

All these pairs focus on the more challenging—and needed—direction, from the higher-resourced language to the shared task languages.

For Question Answering, we use multiple-choice datasets from education and language certification. For Ukrainian, we evaluated on multiple-choice exam questions from the UNLP 2024 Shared Task on LLM Instruction-Tuning for Ukrainian, which is compiled from school graduation examinations on various subjects. For Upper and Lower Sorbian, we evaluated on language certificate exercises which follow the CEFR (Common European Framework of Reference for Languages) scheme.

2.3 Models and Restrictions

We set this shared task in a restricted context with limited resources: The base LLM is fixed to the Qwen 2.5 family (Qwen Team et al., 2025; Yang et al., 2024) and a maximum of 3B parameters. We chose a small model size in order to enable teams with fewer compute resources to participate, and limited models to one family so that results would be more readily comparable. The Qwen 2.5 model family was selected based on zero-shot performance on the Sorbian QA development sets.

Regarding data resources, we suggested training data and provided development sets for all languages (§3). However, the participants were not restricted to the provided datasets. Additional

[1] https://www2.statmt.org/wmt25/translation-task.html

[2] https://glottolog.org

[3] https://www.witaj-sprachzentrum.de

datasets were permitted under the condition that the used resources were open source.

2.4 Phases

The shared task was held in two phases: the development phase started with the release of training and development datasets (when available), and the test phase began when the respective test datasets were made available. The latter also featured a leaderboard on OCELoT where participants could submit their outputs to compare against other teams.

3 Datasets

In the following, we describe the datasets provided for the development and test phases.[4]

3.1 Sorbian Languages

We release the new data for both Sorbian languages under the CC BY-NC-SA licence.

MT For Machine Translation, the WITAJ-Sprachzentrum provided new monolingual and parallel datasets compared to the previous shared task editions for both Sorbian languages. The development dataset is a combination of both old and new sentence pairs. Table 1 lists the number of sentences or sentence pairs with German translations.

The test sets contain 4,000 sentences for each language. The sentence pairs are from the same domain as the training and development datasets.

	Upper Sorbian	Lower Sorbian
parallel train	187,270	171,964
parallel dev	4,000	4,000
parallel test	4,000	4,000
monolingual	47,758 (wiki) 1,071,723 (witaj)	120,501

Table 1: Number of sentences in the newly released mono- and bilingual (paired with German) data for both Upper and Lower Sorbian.

QA The WITAJ-Sprachzentrum organises language examinations for both Upper and Lower Sorbian, from the A1 to C1 levels, according to the CEFR scheme. We use a mix of questions from all five available levels (A1, A2, B1, B2, and C1, from beginner to advanced) for our Question Answering task—more specifically from the reading, grammar, and listening parts. For the listening questions, we rely on the reference transcription of the audio material to convert the exercises; this can lead to comparatively easier questions.

Each language level differs in terms of exercise formats, which can add another level of complexity to the task. While the beginner levels (e.g., A1) often have true-or-false questions on a short text, the advanced exercises (e.g., B2 or C1) typically consist of multiple-choice questions with longer texts and statements, sometimes with up to 16 possible answers.[5]

While the provided development set only contained questions from the A1 to the B2 levels, the test dataset additionally features questions from the C1 certification level. Table 2 shows the number of questions per difficulty level for both data splits.

split	lang	A1	A2	B1	B2	C1	total
dev	hsb	30	28	44	56	0	158
dev	dsb	30	28	44	56	0	158
test	hsb	30	28	44	56	52	210
test	dsb	29	28	44	56	48	205

Table 2: Number of questions in the Sorbian QA datasets per language level.

3.2 Ukrainian

The Ukrainian language track data was sourced from previous editions of MT and QA shared tasks, with the MT test set aligned with this year's General MT competition.

MT For the machine translation subtask, we focused on translation into Ukrainian from two languages: English→Ukrainian (en→uk) and Czech→Ukrainian (cs→uk).

The suggested development data were the combined WMT datasets from the 2022–2024 editions (Kocmi et al., 2022, 2023, 2024). Our test phase was aligned with the WMT 2025 General Machine Translation task for translation into Ukrainian. Parallel data statistics per language pair are shown in Table 3.

The primary difference in dataset sizes comes from the fact that most training and development sets were sentence-based, whereas the test set was

[4]Available at: https://github.com/TUM-NLP/llms-limited-resources2025

[5]More details are available in our shared task repository.

language pair	dev	test
cs→uk	6,263	230
en→uk	5,108	86

Table 3: Datasets statistics per language pair for the Ukrainian MT track.

designed at the paragraph level. This dissimilarity introduced additional challenges for both shared task participants and the evaluation process.

QA We utilised the dataset from the UNLP2024 shared task (Romanyshyn et al., 2024), focusing exclusively on multiple-choice questions. The original data splits were retained, ensuring balanced coverage of all topics across training, development, and test sets, as shown in Table 4.

split	Ukrainian history	Ukrainian lang. and lit.	total
train	910	1,540	2,450
dev	228	385	613
test	348	403	751

Table 4: Datasets statistics per splits and subjects for the Ukrainian QA track.

This dataset comprises machine-readable questions and answers from the Ukrainian External Independent Evaluation (transl. ZNO), the standardized examination required for university admission in Ukraine. It includes exam materials from 2006 to 2023, covering two subjects: History of Ukraine and Ukrainian Language and Literature.

3.3 Relative Difficulty of the Tracks

While we compiled the same task types for both tracks, there are different challenges for our selected languages.

For Ukrainian MT, natural parallel training data is already available, and a variety of both open-source and proprietary translation systems exist. Participants were therefore encouraged to leverage available resources to whatever extent they find appropriate. In contrast, Sorbian resources include parallel and monolingual sentences but are fewer in comparison. NLP tools, more generally, are not available for the two Sorbian languages.

For the QA task, the Sorbian dataset was derived from language certification exams across different levels. Given the languages' low-resource status, the focus is on evaluating whether models can adequately comprehend the language and answer typical document-understanding or grammatical questions. For Ukrainian, the challenge extends further: Models are tasked not only with answering general language questions but also with addressing deeper questions related to Ukrainian literature and history.

4 System Descriptions

Submissions are language-specific; participants could submit to one or more language tracks. However, participants had to submit both the QA and the MT outputs, which had to be generated by the same model.

In addition to our baseline outputs, three teams submitted to the Upper and Lower Sorbian tracks, and one team also submitted to the Ukrainian track. For the final evaluation, each team was asked to choose a primary submission and provide a short description of their approach. Based on these descriptions, we provide an overview of the participating systems here. Table 5 summarises the main characteristics of the three participating primary submissions. For more details, please refer to the respective system description papers.

Baseline (TUM Organisers) Our simple baseline prompts Qwen2.5-3B-Instruct (Qwen Team et al., 2025) with no fine-tuning. The prompts are zero-shot, and an example is shown in Appendix A. We implemented our tasks in the LLM Evaluation Harness framework (Gao et al., 2024) and provided this code to participants for reference.[6]

Team NRC (National Research Council Canada) (Larkin et al., 2025) The NRC submissions focused primarily on the machine translation (MT) component of the shared task. The team explored the impacts of training on Upper Sorbian and Lower Sorbian data separately and together. They also experimented with direct preference optimisation using pairs of correct and incorrect responses to the question answering (QA) set, with limited success. Their submitted systems are based on the Qwen2.5-1.5B-Instruct model, trained using supervised fine-tuning on full model weights using LLaMa-Factory and 4 GPUs (Tesla V100-SXM2-32GB).

[6] https://github.com/TUM-NLP/wmt25-lrsl-evaluation

			NRC	SDKM	TartuNLP
Tracks		hsb & dsb tracks	✓	✓	✓
		uk track	✗	✓	✗
System		Base Qwen model	1.5B	3B	3B
		LoRA	✗	✓	✗
		Quantised	✗	✗	✗
Data	Sorbian MT	Previous WMT MT data	✓	✗	✓
		External data for MT	✗	✓	✓
	MT	Backtranslation	✗	✓	✗
	QA	External data for QA	✗	✓	✗
Training		Joint hsb+dsb training	✓	✗	✓
		Instruction tuning	✗	✗	✓

Table 5: Summary of the three **primary** submissions.

Team SDKM (JGU Mainz) (Saadi et al., 2025) Team SDKM trained three separate Qwen2.5-3B-Instruct models for three languages: Lower Sorbian, Upper Sorbian, and Ukrainian. For Lower Sorbian, the team trained a Qwen2.5-3B-Instruct model by combining both machine translation (MT) and question answering (QA) data. Then, they fine-tuned a joint model on a combined MT and QA dataset. After this fine-tuning, they fine-tuned the model on all the provided QA data and 3k MT data for a second round of fine-tuning. This final fine-tuned model was used for both MT and QA tasks. For Upper Sorbian (hsb), the team followed the same overall approach as with Lower Sorbian. During the QA evaluation of dsb and hsb, they made multiple versions of the same MCQ question with different possible orders and averaged the likelihood of each option, then selected the option with the maximum likelihood. For Ukrainian, they also followed a similar approach, but for QA, they employed retrieval-augmented generation (RAG). The team will make all the data and code they used for pre-processing, training, and their trained model publicly available shortly after the deadline. They used the Qwen2.5-3B-Instruct model as all translation models, for semantic similarity calculation to incorporate retrieval augmented generation.

Team TartuNLP (TartuNLP) (Purason and Fishel, 2025) The TartuNLP system fine-tuned Qwen2.5-3B-Instruct (Qwen Team et al., 2025) on a mixture of monolingual, parallel, and instruction data to support both Lower and Upper Sorbian. The monolingual set included all sentence-level Sorbian data from current and past WMT Shared Tasks, Upper and Lower Sorbian Wikipedia articles (Foundation, 20250520 dump), and Upper and Lower Sorbian documents from Fineweb-2 (Penedo et al., 2025). Parallel data, sourced from WMT (current and past), was reformatted as chat-style instruction pairs, with four epochs of German-to-Sorbian and one epoch of Sorbian-to-German translations. Both monolingual and parallel Sorbian data were repeated four times. The instruction data was collected from Magpie (Xu et al., 2024), Aya (Singh et al., 2024), EuroBlocks (Martins et al., 2025), OpenAssistant (Köpf et al., 2023), and FLAN v2 (Longpre et al., 2023), covering multiple languages. All datasets were deduplicated and packed into 4096-token sequences, with loss applied only to assistant responses in instruction-formatted data. For the final submission, they used beam search with a beam size of 4 for machine translation and one-shot prompting with development set examples for question answering. The final model is published on HuggingFace.[7]

5 Primary Submission Results

5.1 Evaluation Methodology

We evaluate MT with chrF++ (Popović, 2015), computed using SacreBLEU (Post, 2018),[8] and QA with accuracy. ChrF++ ranked slightly higher than BLEU (Papineni et al., 2002) in the WMT 2024 Metrics Shared Task (Freitag et al., 2024). Although neural metrics such as COMET are gen-

[7] https://huggingface.co/tartuNLP/Qwen2.5-3B-Instruct-hsb-dsb

[8] SacreBLEU chrF++ signature: `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.5.1`.

erally known to better correlate with human judgements, we could not consider them as the main ranking criterion because they do not support the Sorbian languages. However, we do consider xCOMET (Guerreiro et al., 2024) for Ukrainian in Section 6.1.

Since our goal is to evaluate the *joint* performance of LLMs on both MT and QA, the ranking in the leaderboard takes into account the scores from all tasks of the track equally. For the final ranking, points are given according to the ranking of the submission in the MT and QA tasks, with the highest-ranked system obtaining the maximum number of points (4 for the Sorbian tracks, 2 for the Ukrainian track). In case of ties, we ranked according to the MT results.

5.2 Leaderboard Results

Tables 6, 7, and 8 show the results of the participating teams' primary submissions for the three tracks: Upper Sorbian, Lower Sorbian, and Ukrainian. The winning team per track and the best submission per task are in bold.

	de→hsb		hsb-QA		final
	chrF++	pts	acc.	pts	pts
TartuNLP	86.33	4	**58.10**	4	8
NRC	**87.20**	4	29.05	1	5
SDKM	75.73	2	55.24	3	5
baseline	13.88	1	42.86	2	3

Table 6: Results for the primary submissions for the **Upper Sorbian** track, ranked by number of points (pts).

	de→dsb		dsb-QA		final
	chrF++	pts	acc.	pts	pts
TartuNLP	78.20	4	**57.56**	4	8
NRC	**78.24**	4	32.20	1	5
SDKM	64.34	2	51.71	3	5
baseline	12.21	1	45.85	2	3

Table 7: Results for the primary submissions for the **Lower Sorbian** track, ranked by number of points (pts).

Upper and Lower Sorbian tracks For both Upper and Lower Sorbian tracks, TartuNLP was the overall winner with high results in both MT and QA. Looking at the tasks of translation and question answering separately, all systems outperformed the baseline for translation, while one system remained below the baseline for question answering. Indeed, if we focus on the translation results only, NRC obtained similar or better performance than TartuNLP, but it showed noticeably lower results for the QA task, since the team chose to focus on MT performance. As they reported, exclusively fine-tuning for MT affects the QA performance negatively, with lower accuracy than the baseline. This confirms our initial assumption and answers our first research question. On the other hand, the NRC submission relies on the smaller 1.5B model and manages to compete with the larger 3B model effectively. This is a promising result for low-resource MT.

For QA, the improvements remain more modest in comparison, with the accuracy increasing by 15 and 11 points, for Upper and Lower Sorbian, respectively. The lack of dedicated training data in the language and the variety of the exercises seem to be the main reasons preventing the models from reaching higher scores.

Ukrainian track For the Ukrainian track (Table 8), only one team (SDKM) participated, achieving results that slightly outperformed the baselines for both MT and QA. In MT, the gains were rather for the closely-related cs→uk pair than for the more distant en→uk pair. Despite accounting for differences in input style between the development and test sets, the results indicate that translating more complex, document-level content remains a substantial challenge for Ukrainian. In QA, the team also surpassed the baseline. For broader comparison, in the UNLP2024 shared task (Romanyshyn et al., 2024), the best-performing model based on Mistral-7B achieved an accuracy of 49, highlighting that smaller models still struggle to reach competitive performance on this task.

6 Deeper Analysis and Discussion

This section focuses on a few analyses of the results. For Ukrainian, to approximate a more advanced evaluation, we compared the participants and the baseline MT results with xCOMET (Guerreiro et al., 2024) (§6.1). For the Sorbian language tracks, we check the QA accuracy per language level (§6.2), contrast the translation performance against other MT approaches (§6.3.1), namely against the current Sorbian-German translator, and perform a manual annotation of translation outputs (§6.3.2).

	cs→uk		en→uk		uk-QA		final
	chrF++	points	chrF++	points	acc.	points	points
SDKM	**8.09**	2	**2.98**	2	**35.82**	2	6
baseline	3.48	1	0.34	1	31.16	1	3

Table 8: Results for the primary submissions for the **Ukrainian** track.

6.1 Evaluating Ukrainian MT with xCOMET

Since the xCOMET metric (Guerreiro et al., 2024) has demonstrated strong performance as a neural-based automatic MT evaluation method and supports Ukrainian, we employed it to gain a deeper insight into the Ukrainian MT results.

We tried both xCOMET-XL[9] and xCOMET-XXL[10] to estimate the difference in significance between the baseline and SDKM team results. The results from xCOMET models are presented in Table 9.

lang. pair	x-wins	y-wins	stat.	p-value
xCOMET-XL				
cs→uk	0.09	0.88	-1.80	0.07
en→uk	0.44	0.48	-0.13	0.89
xCOMET-XXL				
cs→uk	0.01	0.97	-3.17	0.00
en→uk	0.42	0.48	0.04	0.96

Table 9: Comparison of the xCOMET results for the baseline (x) and SDKM (y) **Ukrainian MT** submissions with t-test results.

For the en→uk pair, both XL and XXL models confirmed that the differences between systems were not statistically significant. In contrast, for the cs→uk pair, the XL model results were borderline with respect to the null hypothesis, while the XXL model clearly indicated a significant difference. These findings confirm that the SDKM team achieved better performance for cs→uk translation.

In addition, we conducted a qualitative analysis of the MT outputs at the sample level. A native Ukrainian speaker evaluated the translations for both fluency and adequacy. Representative examples from both models are included in the Appendix B.2. We observe that both models struggled with paragraph-level translation. In several cases, the baseline system failed to generate any output at all. The SDKM team's results, however, are particularly interesting: although their translations did not reproduce the full paragraphs, they captured the main content, resembling a blend of translation and summarisation. This suggests that, with more granular input pre-processing, the model could potentially produce more accurate translations.

[9] https://huggingface.co/Unbabel/XCOMET-XL

[10] https://huggingface.co/Unbabel/XCOMET-XXL

6.2 Detailed Sorbian QA results

model	A1	A2	B1	B2	C1
Upper Sorbian					
TartuNLP	**86.67**	**82.14**	**56.82**	37.50	**51.92**
NRC	50.00	32.14	22.73	19.64	30.77
SDKM	80.00	78.57	**56.82**	**41.07**	42.31
baseline	70.00	57.14	40.91	26.79	38.46
Lower Sorbian					
TartuNLP	**89.66**	71.43	**56.82**	**41.07**	**50.00**
NRC	55.17	42.86	18.18	26.79	31.25
SDKM	82.76	57.14	50.00	37.50	47.92
baseline	65.52	**75.00**	43.18	26.79	41.67

Table 10: Accuracy on the QA datasets per language level for **Upper and Lower Sorbian**.

Table 10 presents the details of the Sorbian QA results for each language level (A1 to C1). We observe that the accuracy drops overall with more difficult question levels for both Upper and Lower Sorbian, except for the B2 level. The lower score at the B2 level for all models compared to the technically more difficult C1 level could be explained by the question types. We recall here that since the questions come from an actual language certification, they are diverse in terms of question type and number of possible answers (cf. §3.1). This also means that we gave equal weight to comparatively more difficult and simpler questions in the main evaluation. If we choose a passing accuracy of 50, most submissions reach a B1 level approximately.

6.3 Detailed Sorbian MT Results

6.3.1 Comparing Sorbian MT

For both Upper and Lower Sorbian, we compare the performance of the submitted systems with existing MT-specific models and quantised versions of LLMs. We present two types of models: the current MT model from the WITAJ-Sprachzentrum (sotra) and a quantised version of the TartuNLP model.

sotra Sotra[11] is the Machine Translation platform developed by the WITAJ-Sprachzentrum since 2019. Dedicated models translate from and to four languages: Upper Sorbian, Lower Sorbian, German, and Czech (except the German-Czech pair), as of 2025. It is based on 800MB fairseq models (Ott et al., 2019), and 50MB quantised versions (INT8 with CTranslate2) have been used for the online version for notably faster outputs. We present the scores for both systems.

Quantised version of the TartuNLP submission Since the sotra website relies on the quantised version for faster inference, we also quantised the model submitted by the TartuNLP team in two different ways. More precisely, we consider a Q4_K_M and a Q8_0 quantised GGUF version of the model.

	de→hsb	de→dsb
NRC	**87.20**	**78.24**
TartuNLP	86.33	78.20
SDKM	75.73	64.34
baseline	13.88	12.21
sotra quantised	79.07	75.92
sotra unquantised	81.52	77.38
TartuNLP Q4_K_M	83.96	75.55
TartuNLP Q8_0	84.83	76.65

Table 11: Comparison of the chrF++ scores on the Upper and Lower Sorbian test dataset for different MT systems.

Results Table 11 presents the MT results on the same test dataset for sotra models and the quantised TartuNLP models. We first observe that the best MT submissions (NRC and TartuNLP) are better than the current Sorbian translator, sotra, for both languages and even with the unquantised version. The gap is larger for Upper Sorbian than Lower Sorbian. We note, however, that the sotra models are older and are thus trained with less data. They are also smaller with around 56M parameters.

Besides, as expected, more aggressive quantisation leads to worse performance. For instance, the Q4_K_M version of the TartuNLP model slightly underperformed compared to the current online sotra model in Lower Sorbian. For Upper Sorbian, the systems still performed better. The inference time is, however, still in favour of the sotra model.

6.3.2 Manual rank annotation of Sorbian MT

As more advanced and reliable automatic metrics are not available for both Sorbian languages, we also evaluate the machine translation outputs through manual rank annotation, despite the high human and time cost associated with it.

Annotation methodology For each Sorbian language, one native speaker ranked the translations from the four systems (including the baseline) for 60 sentences, thanks to the joint organisation with WITAJ-Sprachzentrum. To select the sentences, we first filter the test set (and translations) to avoid cases where two or more systems output the same or too similar sentences (especially for short sentences) or obtain extreme chrF++ scores (e.g., issues with the generation or perfect translation). Then, we randomly select 60 sentences and shuffle the translations before the manual annotation, to reduce bias from the order of the systems.

The two annotators were given the same instructions regarding the ranking. Ranks are assigned to the four translations of the sentence, from 1 for the best translation to 4 for the worst. If two translations are of similar quality, the same score rank can be given (e.g., 1, 2, 2, 4). The reference translation is also given for information purposes. Annotators can additionally put comments beside each machine translation.

Results Table 12 compares the MT system rankings produced according to our main metric (chrF++ score) and the human evaluation for both languages. Unsurprisingly, the baseline model is consistently ranked last by the human annotator for both Sorbian languages; the higher ranks achieved in Lower Sorbian are only due to a large number of ties (e.g., [1, 2, 2, 2]), which blur their poor absolute performance here. The best submitted system is, however, more difficult to conclude; as with the chrF++ score, the annotators also found the NRC and TartuNLP model outputs to be of higher and

[11] https://sotra.app

	rank	chrF++				human			
		NRC	TartuNLP	SDKM	base.	NRC	TartuNLP	SDKM	base.
Upper Sorbian	1st	28	22	10	0	23	26	11	0
	2nd	18	29	13	0	24	25	12	0
	3rd	14	9	37	0	13	9	37	0
	4th	0	0	0	60	0	0	0	60
Lower Sorbian	1st	34	22	4	0	45	37	22	1
	2nd	24	32	4	0	14	18	16	7
	3rd	2	6	52	0	1	5	22	15
	4th	0	0	0	60	0	0	0	37

Table 12: System rankings according to the chrF++ score and the human evaluation for 60 sentences. For instance, the NRC translations in Upper Sorbian were ranked first among the four systems for 28 sentences according to chrF++, while it was 23 times according to the human evaluation.

similar quality. Interestingly, for Upper Sorbian, the latter model seems to be better with the human evaluation.

We also count how often the rankings according to human annotations and chrF++ perfectly match. 27 system rankings (out of 60) are identical for Upper Sorbian and 3 for Lower Sorbian. This difference is due to the higher number of ties given to the systems in the manual annotation of Lower Sorbian.

Qualitatively, the Lower Sorbian annotation comments also showed that the output machine translations still remain unsatisfactory overall, even for the best-ranked system. We present selected examples for both languages in Appendix B.1.

7 Conclusion

The WMT 2025 Shared Task LLMs with Limited Resources for Slavic Languages was the first attempt to evaluate two tasks *jointly*, Machine Translation and Question Answering, and to assess how they impact each other. We focused on three Slavic languages: Upper and Lower Sorbian (paired with German for MT), and Ukrainian (paired with Czech and English). Submissions were constrained to open-source datasets and Qwen 2.5 models below 3B parameters for reproducibility.

Three teams participated. TartuNLP was the overall winner by jointly fine-tuning the model using Sorbian and instruction datasets. NRC won both Sorbian MT tasks with a smaller 1.5B model by focusing on the MT task only. SDKM submitted to all three tracks using additional external datasets as well as data augmentation with machine translation and won on the Ukrainian track. All submissions improved the Machine Translation quality over the baseline Qwen 2.5 3B model. We observe that only focusing on MT negatively affects QA performance, answering our first research question. However, improving the model capabilities on two different tasks remains possible even for small LLMs. This first shared task paves the way for an extension to other tasks and low-resource languages.

Ethics Statement

The shared task focused specifically on smaller models, limiting the submissions to LLMs below the milestone of 3B parameters. This lowers the computational barrier for participants, fostering accessibility and a smaller ecological footprint.

We also strove to have the results reproducible by ensuring that the teams only use both open-source models and datasets. Participants were encouraged to make their model public.

For the Sorbian tracks, the shared task relies on the partnership with the WITAJ-Sprachzentrum. They provided all new Sorbian datasets for both tasks. For the manual evaluation experiment of the Sorbian MT outputs, both annotators were contacted through the WITAJ-Sprachzentrum.

Limitations

The main limitations of this shared task come from the restricted resources. In-domain training data was scarce overall, if not completely lacking, as in the Sorbian QA task, compared to other high-resource languages. Hence, participants needed to resort to data augmentation or external data with machine translation to circumvent this constraint.

Besides, diverging data formatting might have added another layer of complexity, which is orthogonal to our research question. For Ukrainian MT, the sentence-level training data contrasted with the document-level input in the test set, and for Sorbian QA, the exercise variety in type and number of possible answers proved to be a challenge.

Finally, the shared task focused on track-based (i.e., language-specific) approaches for models and not a fully multilingual LLM for all three Slavic languages. The submissions to the Sorbian tracks showed that fine-tuning with both languages proved to be mutually beneficial.

Acknowledgements

This work has received funding from the European Research Council (ERC) under grant agreement No. 101113091 - Data4ML, an ERC Proof of Concept Grant.

We thank the UNLP 2024 Shared Task team: Roman Kyslyi, Mariana Romanyshyn, and Oleksiy Syvokon, for sharing the Ukrainian QA resources.

Moreover, we are grateful for our cooperation with WITAJ-Sprachzentrum for Upper and Lower Sorbian. Beyond the datasets they provided for the previous WMT shared tasks on MT for low-resource languages, this first edition relied on QA data. We thank Tomaš Šołta for providing the language certificate exercises in that regard. We also thank the annotator of Lower Sorbian MT outputs for his time and comments.

Finally, we thank the organisers of the WMT 2025 shared tasks. Firstly, we highly appreciate the help from Martin Popel, Tom Kocmi, Katia Artemova, and Mariya Shmatova for sharing test sets for the Ukrainian MT track with us. We are also especially grateful to Vilém Zouhar and Daniel Deutsch for their advice, Philipp Koehn for helping us with Softconf, and Roman Grundkiewicz for setting up the leaderboard on OCELoT.

References

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Wikimedia Foundation. `https://dumps.wikimedia.org/`.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc.

Samuel Larkin, Chi-kiu Lo, and Rebecca Knowles. 2025. NRC Systems for the WMT2025-LRSL Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical report. *Preprint*, arXiv:2506.04079.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient continual pre-training of LLMs for low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlícek, Vinko Sabolcec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Taido Purason and Mark Fishel. 2025. TartuNLP at WMT25 LLMs with Limited Resources for Slavic Languages Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Hossain Shaikh Saadi, Minh Duc Bui, Mario Sanz-Guerrero, and Katharina von der Wense. 2025. JGU Mainz's Submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm: An efficient state-of-the-art ukrainian llm.

A Baseline Prompts

We show the examples of the MT and QA prompts used in the baseline. Below, we first show the MT prompt for Czech to Ukrainian. The other translation directions match this template, with the language names and ISO-3 codes substituted.

```
Translate the following Czech text to Ukrainian.
Put it in this format <ukr> Ukrainian translation </ukr>.
<cze> {{src_text}} </cze>
```

Below is the QA prompt. The possible answers for Ukrainian are А, Б, В, Г, and Д, so the output of the model is constrained to these 5 options in our evaluation of the baseline. Similarly, the Upper and Lower Sorbian answers are constrained to 1 to 16. While there are not always 16 possible options for each question, we did not observe the baseline model ever choosing a value outside the actual range given in the prompt.

```
{{context}}

Question:
{{question}}

Possible answers:
{{possible_answers}}

Answer:
```

B Illustrative MT Examples per Language Track

We provide several illustrative examples per language track to showcase major successes or problems of the submitted models.

B.1 Upper Sorbian and Lower Sorbian MT Examples

Tables 13 and 14 present two examples of both Upper and Lower Sorbian machine translation. We selected sentences for which notable phenomena are visible; they are, however, not representative of the overall system performance. Hence, we provide the chrF++ score and the translation rank from the human evaluation for information purposes.

The main issue with the baseline translations for both tracks is the language of the output sentence: it frequently features words from the German sentence (sometimes corrupted) or non-existing words in the Sorbian languages or German.

All three submissions improve largely on this issue. We observe a high similarity overall between the translations; if it was easy to identify the baseline output, annotators had to differentiate the other translations sometimes based on how well some nuances were captured.

We also present an output from the SDKM submission in Lower Sorbian (first sentence in Table 14) where non-Latin script characters suddenly appeared. This happened only once in the annotated sentences (i.e., out of 60).

B.2 Ukrainian MT Track Examples

		chrF++	human
de source	Dissoziieren oder sich häufig benebeln und abschalten sind keine günstigen Strategien zur Bewältigung der Situation.		
en translation	*Dissociating or frequently numbing yourself and switching off are not effective strategies for coping with the situation.*		
hsb reference	Disociěrowanje abo so husto zamućić a wotšaltować njejsu přihódne strategije k zmištrowanju situacije.		
baseline	Diszozierun ór sèfórfi ór abefeln ëd abschaltun sàs nesiónn sàs strategyn nòg dżewen.	12.27	4
NRC	Disociěrować abo so husto zamućić a wotšaltować njejsu žane dobre strategije k zmištrowanju situacije.	82.97	2
SDKM	Dissociěrować abo so husto pohłušić a wotpinać njejstej žanej spomóžnej strategiji k zmištrowanju situacije.	56.74	3
TartuNLP	Disociěrować abo so husto zamućić a wotšaltować njejsu žane přihódne strategije k zmištrowanju situacije.	90.06	1
de source	Und paradoxerweise verhalf ihm eben dieser Ultradogmatismus durch alle Spaltungen, Intrigen, Säuberungen hindurch zum Durchbruch.		
en translation	*And paradoxically, it was precisely this ultra-dogmatism that helped him achieve his breakthrough through all the divisions, intrigues, and purges.*		
hsb reference	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě rozpačenja, intrigi, čisćenja k předobyću.		
baseline	Ie ultradogmatizmus verhal iem dëm dazulieb, dëm derselbe dazulieb, dëm derselbe durch alle spaltungen, intrigen, säuberungen hiddur kom.	21.95	4
NRC	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, wučisćenja přez předobyće.	81.23	3
SDKM	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, čisćenja k předobyću.	93.54	1
TartuNLP	A na paradoksne wašnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, čisćenja k předobyću.	87.57	2

Table 13: Examples from the **de→hsb** MT track submissions. We also present the chrF++ scores alongside the human rank annotation.

		chrF++	human
de source	Er nannte als gutes Beispiel die Talsperre Versetal in Nordrhein-Westfalen.		
en translation	*He cited the Versetal dam in North Rhine-Westphalia as a good example.*		
dsb reference	Wón jo pomjenił ako dobry pśikład gaśeński jazor Versetal w Nordrhein-Westfalskej.		
baseline	Er nannte als gutes Beispiel die Talsperre Versetal in Nordrhein-Westfalen.	28.07	4
NRC	Wón jo pomjenił ako dobry pśikład rěcnu zawěru Versetal w Nordrhein-Westfalskej.	78.59	1
SDKM	Wón jo pomjenił ako dobre pśikłady rěcnu zawěru Wortetzer峡在 Nordrhein-Westfalskej.	57.98	2
TartuNLP	Wón jo pomjenił ako dobre pśikłady rěcnu zawěru Versetal w Nordrhein-Westfalskej.	70.57	2
de source	Ob sie wohl jener Mann gesandt hat, dachte Matej und stapfte in den unbekannten Wald.		
en translation	*Matej wondered whether that man had sent him and trudged into the unfamiliar forest.*		
dsb reference	Lěc jo jich ten muski pósłał, jo Matej pómyslił a stupał do njeznateje góle.		
baseline	Matej pohajowat pohadzowat, że gołomany měnka go do niekowiednega węgla wysłał, i chodzil w ten węglu.	19.83	4
NRC	Lěc drje jo jich ten muski pósłał, jo myslił Matej a jo stupał do njeznateje góle.	78.26	1
SDKM	Abo jo wóna togo muskego pósłała, jo se Matej pódał do njeznateje góle.	45.61	3
TartuNLP	Lěc drje jo jich ten muski wupósłał, jo myslił Matej a stupał do njeznatego lěsa.	63.81	2

Table 14: Examples from the **de→dsb** MT track submissions. We also present the chrF++ scores alongside the human rank annotation.

Original	Baseline	SDKM
Vypravěčem spletitého příběhu, točícího se kolem ukradeného velkého diamantu, je neúspěšný boxerský promotér zvaný Turek, který se k divokému kolotoči událostí připletl se svým komplicem Tommym vlastně náhodou. Ukradený diamant má předat kurýr Franky Čtyřprsťák americkému šéfovi Avimu. O drahokam však mají zájem další gangsteři a zlodějíčci, mezi nimi především bezohledný ruský zabiják Boris Břitva, černošští provozovatelé zastavárny Vincent a Solomon a nakonec i nemilosrdný gangster a organizátor nelegálních boxerských zápasů Kruťák...	'[invalid]'	Авторська гангстерська комедія Ґю Річі, з чеською назвою Підфус (1998), базується на попередньому фільмі того ж жанру, Бейсбол, викиньте гроші (1994).
Triptany = specifická léčba akutní migrény. Neuroložka Markéta Tučková a moderní léčebné postupy • mujRozhlas V tuto chvíli vítáme ve studiu ČRo Hradec Králové paní doktorku Markétu Tučkovou z oddělení neurologie Oblastní nemocnice v Náchodě. Dobré dopoledne. Dobrý den. Naše dnešní téma jsou bolesti hlavy, migréna, léčba moderní metody. To všechno bychom dnes chtěli stihnout, i když je to samozřejmě moc. Na úvod tedy, paní doktorko, mohli bychom si rozdělit, udělat představu, jaké jsou dnes možnosti, jak pomoct lidem s migrénou v tom smyslu, jaké léky existují?...	Триптани - це специфічна лікування акутної мигрені. Неврологка Маркета Тучкова і сучасні лікувальні пісокупи • міюРозхлід	Так, тепер профілактика, що є ще однією категорією, є насправді профілактичними препаратами, які не призначені для лікування мігрені, але вони приймаються для зменшення ризику розвитку мігрені, наприклад антигепатичними препаратами, які є моноклональними антитілами проти рецепторів глютену, і це також біологічна терапія мігрені.

Table 15: Examples from **cs→uk** MT track submissions. Both systems have major problems of not being able to capture full paragraph content.

Original	Baseline	SDKM
Start to the season she wanted off and so likewise 83 also off so that brought out another safety car. Christopher parrott involved in that but then great fights for third place especially as costello took the lead. Toth involved in a battle for second, third and fourth. A couple more offs that brought out another safety car. Some racing incidents. Hodson great restart finally to go into second place and that's what nailed that for him but no one was gonna stop michael costello who wins.	Почати сезон вона хотіла і тоді також 83 і офс так щоб піднімнувши ще один сейчар. Кристіпер Парротт був у цьому, але потім великий бой за третье місце особливо як Коостелло викризував перед. Товщина бойу за друге, третье і четверте. Другі пара офсі щоб піднімнувши ще один сейчар. Кількa ринкових подій. Ходсон великий починок навіть до постачання другого місця і це завдає йому це, але жодного не зупинить Майкела Коостелло який перемагає.	Починаючи з сезону, вона хотіла вийти, тож так само і 83, вони обоє були вийняті, що привело до чергового запуску безпеки. Крістофер Парротт був причетний до цього, але потім великі боротьби за третє, четверте і п'яте місце. Дещо аварійних інцидентів. Ходсон був у грі за друге місце, але ніхто не зупинить Майкла Костелло, який виграє.
Didn't expect that if i'm being a hundred percent honest. I didn't expect that i wasn't gonna make a video initially but i put out a tweet saying yay, and there was already a lot of confusion, which is understandable there is so much confusion around what is happening in america right now with the tariffs and there's a couple of points that i want to get to with that ...	'[invalid]'	'[invalid]'

Table 16: Examples from **en→uk** MT track submissions. Both systems have major problems of not being able to capture full paragraph content.