

Findings of the WMT26 Shared Task: Multitask LLMs with Limited Resources

Shu Okabe^{1,2} Marion Di Marco^{*1,2} Kathy Hämmerl^{*1,2} Daryna Dementieva^{1,2}
Lukas Edman^{1,2} Marko Měškank³ Anita Hendrichowa³ Alexander Fraser^{1,2,4}

¹Technical University of Munich (TUM)

²Munich Center for Machine Learning (MCML)

³WITAJ-Sprachzentrum

⁴Munich Data Science Institute

* Equal contribution

Correspondence: shu.okabe@tum.de

Abstract

We present the Findings of the WMT 2026 Shared Task on Multitask LLMs with Limited Resources. This year, we extended the Shared Task to jointly model five diverse NLP tasks, Machine Translation (MT) for five language pairs (eight directions), Question Answering (QA), Spell Checking (SC), Grammar Checking (GC), and Mathematical Reasoning (MR), for two language tracks: Ukrainian and Sorbian (grouping Upper and Lower Sorbian). As the Shared Task is focused on low-resource conditions, the model is restricted to Qwen3.5-2B to remain below the 3B-parameter threshold.

In total, six teams participated across the two language tracks: two in the Ukrainian and four in the Sorbian track. No team participated in both language tracks. All systems have been uploaded to HuggingFace by the participants. Despite our strict restrictions on model and dataset availability, systems were diverse in the resources they used as well as their training and inference strategies. The submitted systems show that a single small model can handle all five tasks, with clear difficulty shown for the challenging MR only.

1 Introduction

Modern Large Language Models (LLMs) have made significant progress in the diversity of tasks they are able to address, but the focus of this progress has been on English and a small number of high-resource languages, where large amounts of data are available. By contrast, performance in lower-resource languages often lags behind or is simply unknown due to a lack of targeted evaluation (Robinson et al., 2023; Ahuja et al., 2023; Dementieva et al., 2026). Additionally, LLMs with many billions of parameters can be difficult to train and inference for members of small language communities, since such efforts typically require significant compute resources.

The challenge for participants in this shared task is to build a model under low-resource conditions to jointly optimise multiple, diverse Natural Language Processing (NLP) tasks. In comparison to last year’s edition of this Shared Task (Okabe et al., 2025), we add three new tasks of varying difficulty: Spell Checking (SC), Grammar Checking (GC), and Mathematical Reasoning (MR). Machine Translation (MT) and Question Answering (QA) were already present in the previous year. Our goal is to encourage the development of stronger models in resource-constrained settings, to provide a structured evaluation of these models, and to study potential synergy or interference effects between the tasks. We constrain submissions to a single base model with 2B total parameters (Qwen3.5-2B).

Several other Shared Tasks provide MT test sets for other low-resource languages or language families, such as Robinson et al. (2025) or Pakray et al. (2025). The Creole MT task additionally solicits data submissions (Robinson et al., 2025). This suggests an ongoing need for low-resource language evaluation and development in MT. Most similar to our Shared Task, the Multilingual Instruction Shared Task (Kocmi et al., 2025a) evaluates on a selection of target tasks in multiple languages, but allows larger models and, this year, leaves out MT evaluation. Meanwhile, the Model Compression task (Gaido et al., 2025) also aims to bring strong performance in compute-constrained settings, but focuses only on the compression approach, evaluating on the General MT test sets.

Our main questions in this shared task are as follows: (i) To what extent can the model handle the five diverse tasks? (ii) Does training for all tasks require dedicated data or training strategies? (iii) How do models based on the more recent but smaller Qwen3.5-2B perform compared to last year?

Although the shared task had strict model and data requirements, this year’s submissions varied

in the resources used, approaches, and training settings. As we released only development and test datasets for the three new tasks (SC, GC, and MR), participants relied on existing related datasets (e.g., for Ukrainian GC), translation of the task-specific data (for MR), or synthetically generated corpora (e.g., for SC). On the training side, we observe that several systems adopt strategies to handle the five tasks without catastrophic forgetting or to preserve capabilities for a subset of tasks. During inference, some participants used a dedicated decoding approach per task. Finally, this year’s base model is a reasoning model, but not all participants enabled the ‘thinking’ mode of the model. Overall, the submitted systems outperformed the zero-shot baseline (Qwen3.5-2B) on the five tasks in most cases, except for MR, which remains a challenge.

The rest of this article is structured as follows: Section 2 describes the target languages, tasks, and restrictions; Section 3 details the released datasets and their evaluation settings; Section 4 relays the submitting teams’ system descriptions. We report test set results in Section 5 and additional analyses in Section 6.

2 Shared task description

2.1 Three languages

Following the WMT 2025 Shared Task on LLMs with Limited Resources for Slavic Languages (Okabe et al., 2025), we continue to focus on the same three languages: the mid-resource Ukrainian and the low-resource Upper Sorbian and Lower Sorbian. In this year’s edition, we split the three languages into two language tracks: the Ukrainian and the Sorbian tracks. This choice stems from the proximity of the two Sorbian languages, as well as the fact that last year’s participants all submitted a joint system for both languages.

Upper Sorbian (ISO code: hsb; Glottocode:¹ uppe1395) and Lower Sorbian (dsb; lowe1385) are minority languages spoken in Germany’s federal states of Saxony and Brandenburg, with 30k and 7k native speakers, respectively. There is an active language community working on the preservation of these languages, notably the WITAJ-Sprachzentrum² (WITAJ Language Center), which also provided parts of the data used in this and previous shared tasks (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022;

Okabe et al., 2025). Nevertheless, the amount of available data for both Sorbian languages still remains limited.

By comparison, Ukrainian (ukr; ukra1253), spoken by approximately 40 million L1 speakers worldwide, is considered a mid-resource language in NLP. Several language-specific pre-training corpora (e.g., Chaplynskyi, 2023) and LLMs (Yukhy-menko et al., 2025) are available.

2.2 Five tasks

This year’s edition of the shared task added three additional tasks, for a total of five tasks to be solved by a single model. In addition to **Machine Translation (MT)** and **Question Answering (QA)**, which follow last year’s shared task setting, we considered **Spell Checking (SC)** and **Grammar Checking (GC)**, which both test aspects of linguistic proficiency. Both also aim to reflect an actual use case from the language community, including language learners. Finally, we deliberately included a very challenging task with **Mathematical Reasoning (MR)**, thereby encouraging the use of the baseline model’s reasoning abilities. Each task is represented by one or more sub-tasks.

2.3 Model and restrictions

We required participants to submit outputs for all five tasks, including all sub-tasks, in their language track with a single model. Regarding the training data, the only restriction was that all data must be publicly available for reproducibility.

The base model for all participants was set to **Qwen3.5-2B** (Qwen Team, 2026) for a fair comparison across submissions and to remain below the 3B-parameter threshold. This model is the closest successor to Qwen2.5-3B (Qwen Team et al., 2025), which was the base model in last year’s edition. Due to timing constraints, we did not run a full comparison study on the development sets to select the model. Still, we considered other models such as LiquidAI/LFM2.5-1.2B-Thinking (Liquid AI, 2026), Qwen3-1.7B (Yang et al., 2025), and Ministral-3-3B-Reasoning-2512 (Liu et al., 2026) in preliminary MT experiments, where they empirically underperformed our chosen baseline model. We also inferenced tencent/HY-MT1.5-1.8B (Zheng et al., 2025) on the Ukrainian track development sets, but found it coped poorly with more diverse prompts and tasks.

The Qwen3.5 model series is trained to support both an instruction mode and a ‘thinking’ mode,

¹<https://glottolog.org>

²<https://www.witaj-sprachzentrum.de>

in which the model produces additional tokens at inference time before answering. However, the developers note that the small 2B model has a greater tendency to enter decoding loops during ‘thinking’. They therefore have it default to instruction mode. In our zero-shot baseline, we decided *against* enabling this reasoning mode, but participants were free to do so. Regardless, we noticed that the model would still produce reasoning tokens before committing to an answer.

3 Datasets

We provide the training, development, and unlabelled test data on our GitHub repository.³ The licence for each material is detailed in the repository and differs per task within a language track. We note that we only release development and test sets for the three new tasks.

3.1 Machine Translation

For the Ukrainian track We retest on the same dataset as last year’s edition. This dataset was aligned with the General MT task at WMT2025 (Kocmi et al., 2025b) and proved to be very challenging for our resource-constrained setting last year. The test set for Ukrainian includes two translation directions: en→uk and cs→uk. Table 1 reports the four released dataset sizes. This task features two sub-tasks. We note that the test set contains paragraphs (in contrast to the dev sets with sentences).

lang. pair	cs→uk	en→uk
dev	6,263	5,108
test	230	87

Table 1: Number of translation instances for the **Ukrainian MT** task (same sets as in our 2025 edition).

For the Sorbian track As in previous editions, the WITAJ-Sprachzentrum provided new training (monolingual and parallel), development, and test datasets. Further, we consider additional language pairs, for a total of six translation directions: de→hsb and de→dsb as in last year’s task, their reverse directions hsb→de and dsb→de, as well as translation between the two Sorbian languages: hsb→dsb and dsb→hsb. Regarding test datasets, we use the 2025 Shared Task edition split

for de→hsb and de→dsb and add the smaller but newly provided splits. For the four other directions, the test sets are new and do not overlap.

In addition to providing new test data, we also retest on last year’s machine translation test sets for comparison. Unfortunately, a subset (777 sentences) of last year’s de→dsb test set was accidentally included in this year’s official training data, making this specific comparison less valuable. In our final scores, we therefore evaluate the sentences *without* the overlap for a fair assessment. The new test sets were not affected by this issue.

Table 2 reports the number of sentences in all the released sets for the MT task of the Sorbian track. This task has six sub-tasks.

Upper Sorbian (hsb)		Lower Sorbian (dsb)			
mono.	512,671	38,028			
de→hsb		de→dsb		hsb→dsb	
→	←	→	←	→	←
train	23,116	30,560		67,845	
dev	4,000	4,000		4,000	
test	4,822 4,000	5,352* 4,000	4,000 4,000		

Table 2: Number of sentences in this year’s released monolingual and parallel data for both Upper and Lower Sorbian. * We report the full test dataset size for de→dsb; removing the overlapping sentence pairs leads to the 4,575 sentences used for our final evaluation.

3.2 Question Answering

For the Ukrainian track As in the previous edition of our shared task, we use the UNLP2024 dataset (Romanyshyn et al., 2024), focusing exclusively on multiple-choice questions. The dataset contains machine-readable questions and answers from Ukraine’s standardised university admission examination (ZNO), covering exam materials from 2006 to 2023 in two subjects: History of Ukraine and Ukrainian Language and Literature.

This year, we additionally include the Ukrainian MMLU dataset introduced by Yukhymenko et al. (2026). It was constructed through rigorously evaluated automatic translation of the original MMLU dataset (Hendrycks et al., 2021a), which covers 57 tasks spanning domains such as elementary mathematics, history, computer science, and law. Achieving high performance on MMLU, therefore, requires models to demonstrate broad world knowledge and strong problem-solving abilities.

Table 3 summarises the splits of the two datasets for the Ukrainian Question Answering task. This

³<https://github.com/TUM-NLP/llms-limited-resources2026>

task has two sub-tasks (ZNO and MMLU).

set	train	dev	test
ZNO	2,450	613	751
MMLU	1,531	285	14,042

Table 3: Number of questions for the two sub-tasks of **Ukrainian QA**. The ZNO dataset was also used in the 2025 edition.

For the Sorbian track We use the same QA data for both Sorbian languages as in last year’s edition. It is based on actual questions from language certificates with 5 proficiency levels (A1, A2, B1, B2, and C1; from beginner to advanced) measured according to the CEFR scheme. The dataset was shown to be hard for intermediate and advanced levels due to its variety of tasks (e.g., reading comprehension of long texts, grammar questions) and of formatting (number of answers ranging from 2 to 16).

Table 4 presents the number of questions for the Sorbian QA track. There are two sub-tasks for this task in the Sorbian track (hsb and dsb). For the development phase, we only released questions from the A1 to B2 levels, while for the test phase, we additionally feature harder C1-level questions.

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 4: Number of questions in the **Sorbian QA** datasets per language level.

We also have a hidden QA dataset of 500 questions. Due to logistical constraints, not all systems could be evaluated, leaving this assessment for future editions.

3.3 Spell Checking (both tracks)

The spell checking task aims to assess whether models can identify and correct a character-level error in a sentence. It mainly aims to represent typos rather than grammatical errors, which are tested in a dedicated task (§3.4).

We standardise the task as follows: given a sentence that may contain an error in one of its words,

wrong input:	30.000 opozicionelnych bcuhu zajeći
outputs	detection: bcuhu correction: buchu
correct input	30.000 opozicionelnych buchu zajeći
outputs	detection: CORRECT correction: CORRECT

Figure 1: Example of input and outputs for the **spell-checking** task in Upper Sorbian (in English, ‘30,000 opposition members were arrested’). The wrong instance shows a character swapping transformation.

the model must predict which word is wrong (detection) and correct it (correction). Figure 1 illustrates the task for both wrong and correct sentences.

We created two *evaluation* datasets of 2,000 sentences each for the development and test splits of the spell-checking task. Both were balanced, where half of the sentences were correct, with no errors to spot or change, and half were synthetically corrupted, with a single modified word.

Sentence source We use new sentences from the parallel splits originally provided for MT by the WITAJ-Sprachzentrum for Upper and Lower Sorbian. They do not overlap with the (previously or newly) released MT datasets (§3.1). For Ukrainian, we combined formal data from Dementieva et al. (2025),⁴ court decisions,⁵ and job descriptions (Drushchak and Romanyshyn, 2024).⁶ Table 5 displays the dataset statistics for the spell-checking task. The Ukrainian track features one sub-task, while the Sorbian track has two.

We consider four types of modifications: character deletion, insertion, substitution, and swapping. We further categorise substitutions into *random*, where any character can replace another, and *subtle* to replicate errors which seem more ‘natural’, such as phonetically similar character pairs. All errors affect at most two characters. For each of the five modification categories, we assigned 20% of the corrupted sentences (i.e., 200 sentences) for equal representation.

3.4 Grammar Checking (both tracks)

The goal of this task is to identify a grammatical mistake in a sentence, with one incorrect word per corrupted sentence. We created development and

⁴<https://huggingface.co/datasets/ukr-detect/ukr-formality-dataset-seminatural>

⁵<https://huggingface.co/datasets/lang-uk/court-decisions-uk>

⁶<https://huggingface.co/datasets/lang-uk/recruitment-dataset-job-descriptions-ukrainian>

	train	dev	test
uk, hsb, dsb	-	2,000	2,000

Table 5: Number of **Spell Checking** sentences per language.

test sets with 2,000 sentences (Ukrainian, Upper Sorbian) and 1,250 (dev) and 1,726 (test) for Lower Sorbian. Each dataset is balanced with 50% corrupted and 50% correct sentences. In the corrupted sentences, one word was replaced with a different inflected form of the same lemma. With the corrupted words being valid word forms, this task aims at evaluating the ability to rate a word’s correctness in a sentence context, rather than identifying incorrect spelling. In addition to identifying the incorrect word, the model also needed to predict the correct word.

The words are corrupted at the level of grammatical case, for example, by replacing a word form in dative case (Dat) by a word form of the same lemma in accusative case (Acc); all other features (such as number (singular (Sg) or plural (Pl)) and gender) remain the same. As a basis for selecting and replacing words, we used UNIMORPH⁷ (Batsuren et al., 2022) for Ukrainian and Lower Sorbian and soblexx⁸ for Upper Sorbian, where no UNIMORPH data is available. Both resources provide inflected word forms and their morphological features paired with the lemma.

Since many word forms can have different analyses, depending on the context, we need to ensure that the generated corrupted sentences are indeed grammatically incorrect. For this reason, we restrict words to be corrupted to those with only one morphological analysis. Corrupted words can be nouns and adjectives: for nouns, we change grammatical case in order to create a mismatch to the expected realisation and potential agreement errors. Additionally, we allow for replacing number and gender in adjectives to create agreement errors.

We also made sure that the inflected form selected as replacement in the corrupted sentence does not share features with the original word form, which could lead to an accidentally correct sentence with a different meaning. For example, if for an original noun (Acc.Sg), a selected form (Dat.Sg) can also have the analysis Acc.Pl, the generated

corruption might still be grammatically correct (assuming there are no further words in the phrase resulting in an agreement error), but with a change in meaning (Pl vs. Sg).

Finally, we used gemini-3-flash-preview to check the generated datasets by asking the model whether the sentence is correct or contains an error, keeping only those sentences in the test set where no errors were detected for presumably correct sentences, and the target word was identified as incorrect in corrupted sentences.

We sample sentences from the same dataset as in the spell-checking task. We note that the sentences do not overlap between the two tasks. Table 6 summarises the three datasets for the grammar checking task. Here also, the Ukrainian track features one sub-task, while the Sorbian track has two. Figure 2 illustrates the task, while Figure 9 in Appendix B shows the details for an example instance.

wrong input:		Wusypaj pššk z crjejam!
outputs	detection:	crjejam
	correction:	crjejow
correct input:		Wusypaj pššk z crjejow!
outputs	detection:	CORRECT
	correction:	CORRECT

Figure 2: Example of input and outputs for the **grammar-checking** task in Lower Sorbian (in English, ‘Shake the sand out of your shoes!’). We show the same sentence in two versions here for illustration purposes; no sentence is ‘duplicated’ in the released sets.

lang	uk	hsb	dsb
train	-	-	-
dev	2,000	2,000	1,250
test	2,000	2,000	1,726

Table 6: Number of **Grammar Checking** sentences per language.

3.5 Mathematical Reasoning (both tracks)

Input: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?
Output: 540

Figure 3: Example of input and output for the **maths reasoning** task. The example is in English and is drawn from PolyMath (Wang et al., 2025), the original source for PluraMath (Dementieva et al., 2026).

⁷<https://unimorph.github.io/>

⁸<https://soblexx.de/>

The fifth and last task is mathematical reasoning, where a mathematical exercise has to be solved, as illustrated in Figure 3. For this task, we use the corresponding set of 250 instances per language from PluraMath (Dementieva et al., 2026). PluraMath features an extended version of PolyMath (Wang et al., 2025) for 18 more languages—including Ukrainian, Upper Sorbian, and Lower Sorbian—and consists of four task difficulty levels: low, medium, high, and top.

The extension to new languages was obtained through a human-in-the-loop pipeline. First, initial translations were generated using the best-performing automatic translation system for each language. Then, one or two native speakers reviewed the translated data and refined the translations. This process helps ensure the high quality of the PolyMath dataset extension.

Given that the high and top levels contain challenging Olympiad-level problems, we focus on the two lowest levels, low and medium, under our constrained shared task setup. For the development phase, we release the first 12 instances of both levels (i.e., 24 instances in total). For the test set, all 250 exercises per language are released. We explicitly asked participants not to use the overlapping development set for training, nor the original PolyMath dataset in any existing language.

Table 7 summarises statistics of the datasets released for the MR task. The Ukrainian track features one sub-task, while the Sorbian track has two.

	train	dev	test
uk, hsb, dsb	-	24*	250

Table 7: Number of **Maths Reasoning** instances per language. * the development set is *included* in the final test set.

4 System descriptions

Three teams submitted to the Ukrainian track and four submitted to the Sorbian track. One team submitted outputs to the Ukrainian track but did not submit a system description. We report their results for reference purposes but do not consider them as shared task participants. We note that two teams (Zolint, TUM-HN) feature a TUM affiliation, one of which includes a Shared Task organiser as an advisor. These teams did not receive any additional information or special treatment.

Participants were asked to submit their model outputs on OCELoT for the leaderboard and to publish their models, preferably publicly on HuggingFace. All teams which sent a system description followed this recommendation. Since a team is allowed to submit multiple times, we only evaluate their selected *primary* submission. Below, we report the short descriptions of their system that each participant was asked to provide, with minor edits for layout and flow purposes.

4.1 Zero-shot baseline (TUM Organisers)

We evaluate the Qwen3.5-2B baseline model in a zero-shot fashion, without using explicit reasoning tokens for both tracks. Question answering is inferred in a forced-decoding manner, where only the allowable answer tokens are considered, while all other tasks are inferred through open-ended decoding. For the generative tasks, we set our decoding parameters according to the recommendations of the model developer and apply the chat template, with the model’s reasoning token budget set to zero. The maximum token budget depends on the task, with just 64 tokens for GC and SC compared to 512 tokens for MR. Our baseline prompts are listed in Appendix A. We implemented all tasks in the LLM Evaluation Harness (Gao et al., 2024) and provided this code to participants for reference, along with a script to convert outputs from the framework to the final submission format.

4.2 Ukrainian track

Zolint (UTokyo & TUM) (Yang et al., 2026)

This team submitted a LoRA-based multitask fine-tuned Qwen3.5-2B-Instruct model for the WMT 2026 Shared Task on LLMs with Limited Resources.⁹ The system jointly addresses English–Ukrainian and Czech–Ukrainian machine translation, question answering, spelling correction, grammar correction, and mathematical reasoning. Training data include the official shared-task resources together with publicly available datasets such as WMT 2025 parallel corpora, NLLB (Team et al., 2022), UA-GEC (Syvokon et al., 2023), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), and additional synthetic data for spelling correction and question answering. They adopt a two-stage instruction-tuning strategy: the first stage focuses on translation and correction tasks to strengthen Ukrainian generation, while the

⁹<https://huggingface.co/MDK-YLC/Qwen3.5-2B-WMT26-Ukrainian-LoRA>

Tracks			Ukrainian		Sorbian			
			Zolint	TUM-HN	HeyBusan	HITSZ	HSE	LT3
Resources	Sorbian MT	Previous (2020–2022) [†] WMT MT data	✗	✗	✓	✓	✗	✓
		2025 WMT MT data	✗	✗	✗	✓	✗	✗
		External monolingual data for Sorbian MT	✗	✗	✓	✗	✗	✓
		Back-translation	✓	✗	✗	✗	✗	✓
	QA	External data for QA (existing)	✓	✓	✗	✗	✗	✗
		Translating data for QA	✓	✗	✗	✗	✗	✗
		Synthetic data generation for QA	✗	✗	✗	✓	✓	✓
	SC	Use of external task-related dataset	✗	✓	✗	✗	✗	✗
		Synthetic data generation	✓	✗	✓	✓	✓	✓
		Use of dictionary	✗	✗	✓	✗	✗	✓
	GC	Use of external task-related dataset	✓	✓	✗	✗	✗	✗
		Synthetic data generation	✗	✗	✓	✓	✓	✓
		Use of linguistic resources*	✗	✗	✓	✗	✓	✓
	MR	External MR data (existing, as is)	✗	✓	✗	✓	✓	✗
		External MR data (translated)	✓	✗	✓	✓	✗	✓
		MR data language	UK	EN	DE/EN	EN/HSB/DSB	EN	HSB/DSB
	Instruction data in Sorbian (translated)		✗	✗	✗	✓	✗	✗
Method	Multi-step training (with different settings)		✓	✗	✓	✓	✓	✓
	Training step for a subset of tasks		✓	✗	✗	✗	✗	✓
	Weight merging		✗	✗	✓	✗	✗	✗
	(Q)LoRA		✓	✓	✗	✗	✓	✓
Task-specific	QA	Strategy to reduce position bias	✓	✗	✓	✗	✗	✓
	MT	Retrieval augmentation	✗	✗	✗	✗	✗	✓
	Tasks	Majority voting	✗	✗	✓	✗	✗	✗
	MR	Thinking	✓	✗	✗	✓	✗	✗
	Other tasks	Thinking	✓	✗	✗	✗	✗	✗
GPU	Training or inference		-	H100	RTX 3090	8 GPUs	2 RTX 3090	H200
LLM use	For data generation (including translation)		✗	✗	Qwen3.5-2B	Qwen3.7-Max, gpt-5.5	✗	Qwen3.5-2B (FT)
Other models	For translation		✓	✗	✗	✗	✗	✗

Table 8: Summary of the six **primary** submissions to our shared task: 2 systems for the Ukrainian track and 4 for the Sorbian track. Information is compiled based on their respective system description. [†] We group here all previous datasets released between 2020 and 2022; however, not all teams have used the whole set. * We denote the following items used as linguistic resources for GC: a dictionary, a reference corpus, or a morphological dataset.

second stage introduces reasoning and question answering with replayed first-stage data to mitigate catastrophic forgetting. All tasks are unified into an instruction-following format and optimized using a shared LoRA adapter.

TUM-HN (Temesgen et al., 2026) This team fine-tuned the Qwen3.5-2B model for Ukrainian in a multitask learning setting, covering Machine Translation (MT), Question Answering (QA), Spell Checking (SC), Grammar Checking (GC), and Maths Reasoning (MR).¹⁰ For MT, they sourced English–Ukrainian and Czech–Ukrainian data from Tatoeba (Tiedemann, 2020). For QA, they used data from ZNO (Romanyshyn et al., 2024), MMLU (Yukhymenko et al., 2026), and UA-SQuAD (Ivanyuk-Skulskiy et al., 2021). For MR, they used the OlympiadBench (He et al., 2024) dataset, restricting it to the mathematics and physics domains, and they also included the MathMist dataset (E Sobhani et al., 2026). For SC

and GC, they used the Grammatical Error Correction and Fluency Corpus for the Ukrainian Language (UA-GEC; Syvokon et al., 2023) dataset. All datasets were in Ukrainian, except for the MR datasets, which were available only in English.

All datasets were reformatted as chat-style instruction pairs. The model was first fine-tuned for 3 epochs, then further fine-tuned for up to 10 additional epochs with early stopping (patience of 3). Loss was computed only on the assistant responses. They inferred the model in the zero-shot setting over the test set.

4.3 Sorbian track

HeyBusan (Khasykov, 2026) This team performed a full-parameter fine-tune of Qwen3.5-2B for all five Sorbian-track tasks,¹¹ merged 0.9/0.1 in weight space with the base model to restore maths/QA ability lost to multitask Supervised Fine-Tuning (SFT).

¹⁰<https://huggingface.co/Tsedeniya/qwen3.5-2B-sft-wmt2026-CFT>

¹¹<https://huggingface.co/mkhasykov/qwen3.5-2b-wmt26-sorbian>

Training mixes the official task data (including a final sealed pass over the full dev sets; MR dev never used for PolyMath compliance) with decontaminated WMT20–22 de-hsb/de-dsb/hsb-dsb corpora, FineWeb-2 Sorbian monolingual text, a public instruction mix as forgetting control, lexicon-derived synthetic spell/grammar-checking data (soblex/dolnoserbski full-form lexicons: paradigm-swap GC errors, spell-checker-rejected SC corruptions, 50% correct hard negatives), and STaR-style (Zelikman et al., 2022) verified German reasoning chains generated by their own model on German translations of GSM8K-train.

The team relies on per-task decoding: with beam-4 for MT, self-consistency voting for SC and GC (maj@16, with votes filtered by the full-form lexicons: corrections that are not attested word forms are dropped) and MR (16 samples per view over the original question and its German translation), and for QA log-likelihood answer selection averaged over all cyclic rotations of the answer-option order (position debias).

HITSZ (Zhang et al., 2026) Their system is based on Qwen3.5-2B.¹² They first performed mixed continual pre-training on Upper Sorbian, Lower Sorbian, German, and English monolingual and parallel corpora, followed by full-parameter multitask supervised fine-tuning for machine translation, question answering, spelling correction, grammar correction, and mathematical reasoning. Task-specific prompts were aligned with the official evaluation format, and native thinking mode was used only for mathematical reasoning.

HSE (Diskin, 2026) This team submitted a single Sorbian multitask system based on Qwen/Qwen3.5-2B, adapted with an all-linear rank-16 LoRA and merged into one BF16 model for inference.¹³ The model was trained with weighted token-level cross-entropy on a 278,488-example mixture covering machine translation, question answering, spelling correction, grammar correction, and mathematical reasoning. The mixture contains 234,594 examples from the released WMT2026 Sorbian parallel MT data; 15,000 synthetic spelling-correction, 1,800 controlled grammar-correction, and 20,000 multiple-choice QA proxy examples derived exclusively from the released MT training corpus;

¹²<https://huggingface.co/worthing-zhang/sorbian-sft-v3-ck3250>

¹³https://huggingface.co/yhn112/WMT_Limited_Resources_Sorbian

and 7,094 answer-only examples from the public English GSM8K training split. Source-group-disjoint splits and collision filtering were used to prevent overlap with official development and test data. After the initial training, the team applied one conservative optimizer-reset continuation epoch on the same mixture at a lower learning rate and selected checkpoint 8703 using a pre-registered multi-task rule on the official Sorbian development tasks. The same merged checkpoint, the unchanged official prompts, and converter were used for all five OCELoT files.

LT3 (Moerman and Tezcan, 2026) A single QLoRA-tuned Qwen 3.5 2B serves all five Sorbian tasks.¹⁴ Training combines the organiser data with public German–Sorbian parallel corpora (WMT20-22), back-translation of monolingual Sorbian with their own model, and synthetic task data. This synthetic data covers typo injection for spell-checking, morphological paradigm-swap corruptions (soblex and dsb-spell) for grammar-checking, machine-translated public maths seeds (GSM8K and MATH, with PolyMath excluded) for maths reasoning, and cloze-style QA. The system uses retrieval-augmented inference. Half of the training examples carry a fuzzy-retrieved exemplar, and at inference, MT retrieves three exemplars from a real plus back-translated translation memory that is fully enumerated, hashed, and deduplicated against all test sources, with no web access. Grammar-checking consults corpus bigram-attestation statistics additionally from the same enumerated corpora. The frozen model plays two roles. For MT and MR, a generator produces the output text. For QA, SC, and GC, it is a scorer that ranks candidate answers or corrections by the likelihood it assigns them. On top of the frozen model, several inference strategies are employed. Grammar gap-fill QA items are answered by scoring each option in place with the model’s own log-likelihood, and shared option banks are solved as a global assignment. Grammar-checking fuses the model’s detection with a full-paradigm candidate scorer and a corpus bigram-attestation witness. Spell-checking constrains corrections to dictionary neighbourhoods and adds a frequency-guarded diacritic detector. Maths uses a retry-on-empty ladder with answer-notation formatting. The primary system folds the development sets into training.

¹⁴<https://huggingface.co/collections/TomMoeras/wmt2026-multitasksorbiannlp-lt3>

	MT (chrF++)	pts	QA	pts	SC	pts	GC	pts	MR	pts	final points
Zolint	49.86	4	41.15	3	66.43	4	45.18	4	21.60	4	19
koshi*	16.80	2	46.99	4	52.50	3	30.55	3	0.40	1	13
baseline	18.46	3	36.33	1	8.38	1	2.18	1	20.00	3	9
TUM-HN	16.08	1	40.13	2	16.50	2	6.80	2	4.40	2	9

Table 9: **Ukrainian track** results summary. Within each task, subtasks are averaged with equal weight. Points are assigned as the inverse of score rank. The best task scores and winning teams are in **bold**. * koshi submitted a system output but did not provide a system description; we report their results for information purposes.

	MT (chrF++)	pts	QA	pts	SC	pts	GC	pts	MR	pts	final points
HeyBusan	72.75	5	55.92	3	78.03	5	79.52	5	29.00	4	22
LT3	69.57	4	65.05	5	76.11	4	78.32	4	30.40	5	22
HITSZ	63.12	3	57.33	4	73.13	3	62.11	3	18.60	3	16
HSE	61.50	2	46.77	2	63.63	2	50.02	2	5.00	1	9
baseline	21.68	1	43.39	1	6.66	1	1.34	1	5.40	2	6

Table 10: **Sorbian track** results summary. Within each task, subtasks are averaged with equal weight. Points are assigned as the inverse of score rank. The best task scores and winning teams are in **bold**.

4.4 Summary

Table 8 summarises the main characteristics of the six primary submissions. On the resource side, due to data scarcity or the absence of released dedicated training data for some tasks, external corpora were used, and synthetic data were generated. Lexical resources were also used to help SC and GC. For MR, existing corpora in high-resource languages (such as English) were used as is or translated into the target language or a relevant high-resource language (e.g., German for the Sorbian track). Participants opted for full-parameter training or parameter-efficient fine-tuning with LoRA (Hu et al., 2022) or QLoRA (Dettmers et al., 2023). Dedicated inference strategies were also used for better prediction performance. We also note that QA, SC, and GC were solved both using generation- and scoring-based approaches. Finally, not all participating teams used the ‘thinking’ mode of the base model.

5 Primary submission results

5.1 Evaluation method

5.1.1 Task evaluation

To help participants with the evaluation of their models, we provided an evaluation tool from the lm-evaluation-harness again this year.¹⁵

¹⁵<https://github.com/TUM-NLP/llms-lim-res-eval-2026/>

Machine Translation For a unified MT evaluation, we use BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015), computed with sacreBLEU (Post, 2018).¹⁶ chrF++ remains our main MT metric for consistency across translation directions.

Question Answering We compute the standard accuracy for QA as in the previous edition. Since both tracks feature two sub-tasks (i.e., dataset language or sources), we give them equal weight by reporting their mean sub-task accuracy, despite the difference in size for the Ukrainian track.

Spell Checking and Grammar Checking We evaluate the two tasks according to two criteria: error *detection* (i.e., did the model identify the erroneous input?) and error *correction* (i.e., did the model modify the error in the correct form?). We compute the standard accuracy for both fields.

Maths Reasoning We evaluate the maths reasoning task using the standard exact-match accuracy. Instead of a pure string-level match, we use MathVerify¹⁷ to handle non-identical but equivalent maths formulae (e.g., $\frac{1}{2}$ and 0.5) to have a better assessment of the models.

¹⁶SacreBLEU BLEU signature: nrefs:1 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.6.0. SacreBLEU chrF++ signature: nrefs:1 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.6.0.

¹⁷<https://github.com/huggingface/Math-Verify>

5.1.2 Final result computation

For each track task (e.g., Ukrainian QA) with several sub-tasks (e.g., ZNO and MMLU), we average the scores per sub-task equally. We follow last year’s point system to score each participating team. We assign the maximum score (the number of evaluated systems; 4 for Ukrainian, 5 for Sorbian) to the best performing system, down to 1 point per task. The final ranking is the sum of all points. Ties between systems are allowed.

5.2 Ukrainian track

We summarise per-task results and the resulting points for the Ukrainian track in Table 9. Detailed results, with per-language-pair metrics for MT and the values for the wrong vs. corrected words in SC and GC, are reported in Table 18 in Appendix C.

The Zolint team won this track, reaching the best score across four of the tasks (all except QA). It also clearly outperforms the baseline overall. We observe that this system largely outperforms the others for Machine Translation. Interestingly, the baseline achieves 20 points of accuracy for MR, while the koshi and TUM-HN systems struggle with it. Finally, we note here that on the same test datasets (MT and QA-ZNO), Zolint outperforms last year’s best system.

5.3 Sorbian track

We summarise per-task results and the resulting points for the Sorbian track in Table 10. We report detailed results in Table 19 in Appendix C with per-language scores, BLEU as an additional metric for MT, and the comparison for wrong-word and corrected-word matches in SC and GC.

For the Sorbian track, the HeyBusan and LT3 teams are tied winners, both with 22 total points. All submitted systems outperform the baseline overall, with notable improvement for the MT, SC, and GC tasks. It is worth noting that both winning systems obtain rather similar performance on all tasks except QA, despite clear differences in their systems in terms of data, training (full fine-tuning and (Q)LoRA), or inference strategies (cf. Table 8).

6 Complementary analyses

We perform six additional analyses to better understand the model outputs. Section 6.1 computes xCOMET (Guerreiro et al., 2024) scores for MT evaluation when the target language is supported (Ukrainian and German, in our case). Section 6.2

compares this year’s Sorbian MT systems with the current state of the art and reports the results of the manual ranking of a subset of translations. Section 6.3 compiles the Sorbian QA accuracy scores per language certification level. Sections 6.4 and 6.5 analyse the accuracy distribution according to the SC modification type or part of speech for GC, respectively. Section 6.6 analyses the outputs (including reasoning traces) for MR. We also report the system rankings after statistical tests per task in Appendix D.

Team	cs→uk	rank	en→uk	rank
Zolint	21.38	1	19.42	1
TUM-HN	20.34	1	16.78	2
koshi	18.16	3	15.55	2
baseline	17.13	3	12.29	4

Table 11: Mean xCOMET scores for **Ukrainian MT**. The rank columns give the significance-cluster rank per direction (paired t -tests, $p < 0.05$).

Team	hsb→de	rank	dsb→de	rank
HeyBusan	68.78	1	60.54	1
LT3	66.42	2	59.04	2
HITSZ	62.18	3	55.26	3
HSE	53.08	4	51.03	4
baseline	21.03	5	16.59	5

Table 12: Mean xCOMET scores for Sorbian-to-German translation (part of the **Sorbian MT** track). The rank columns give the significance-cluster rank per direction (paired t -tests, $p < 0.05$).

6.1 xCOMET for supported languages

For translation into Ukrainian (cs→uk and en→uk) and into German (hsb→de and dsb→de), we compute reference-based xCOMET (Guerreiro et al., 2024) in addition to chrF++ and BLEU. This learned metric generally correlates better with human judgements than surface-level metrics, but it does not support the two Sorbian languages. We compute per-sentence scores with xCOMET-XXL, then run bootstrap resampling and pairwise t -tests¹⁸ to determine significance clusters.

Tables 11 and 12 report mean scores, along with ranks per translation direction. For the Ukrainian track, the results show that overall MT performance for the language remains relatively poor,

¹⁸As implemented by the COMET CLI.

with Zolint still in the lead but by less than the surface-level metrics suggest. TUM-HN pulls slightly ahead of koshi under this metric. For translation from Sorbian into German, the ranking from xCOMET-XXL agrees with the surface-level metrics. Overall quality is rated much higher here, since this dataset is comparatively simple.

6.2 Sorbian MT analyses

6.2.1 Comparison of Sorbian translation

We compare this year’s participating systems with last year’s winner (TartuNLP) and the Sorbian reference translator, *sotra*, for *all six* translation directions.

TartuNLP The TartuNLP system¹⁹ (Purason and Fishel, 2025) is based on Qwen2.5 (Qwen Team et al., 2025) with 3B parameters and was officially evaluated on the test set for the two *de*→*hsb* and *de*→*dsb* directions last year. We output the translations using our evaluation harness.

sotra In contrast, *sotra* features dedicated models trained for each of the six directions. It has been developed by the WITAJ-Sprachzentrum since 2019 and is available online.²⁰ We report here the INT8 quantised *sotra* models of 56M parameters (50MB; *ctranslate2*).

	<i>de</i> → <i>hsb</i>	<i>de</i> → <i>dsb</i>	<i>hsb</i> → <i>de</i>	<i>dsb</i> → <i>de</i>	<i>hsb</i> → <i>dsb</i>	<i>dsb</i> → <i>hsb</i>
TartuNLP	83.81	73.29	76.65	75.64	85.05	84.91
<i>sotra</i>	81.96	72.54	72.70	78.40	84.33	84.23
HeyBusan	70.26	63.06	69.43	69.42	81.96	82.38
LT3	64.55	58.90	67.19	66.56	80.12	80.09
HITSZ	60.04	52.37	62.42	60.92	72.35	70.62
HSE	56.27	52.72	52.70	54.61	77.19	75.50
baseline	13.67	10.42	27.05	23.18	27.10	28.66

Table 13: Comparison of Sorbian machine translation performance (chrF++) with last year’s winning system (TartuNLP) and the dedicated *sotra* translation.

Table 13 displays the chrF++ scores for all six translation directions. First, we observe that the best system on MT from this year’s edition is behind both TartuNLP’s model and *sotra*. The gap is wider for pairs involving German than between the two Sorbian languages. Therefore, the state-of-the-art open MT system for Sorbian translation from our Shared Task series remains TartuNLP. We note that its training featured all six language directions.

¹⁹<https://huggingface.co/tartuNLP/Qwen2.5-3B-Instruct-hsb-dsb>

²⁰<https://sotra.app/>

6.2.2 Manual annotation for Sorbian MT

Thanks to the WITAJ-Sprachzentrum, one native speaker manually annotated a subset of MT system outputs for each Sorbian language. Following last year’s protocol, the two annotators *ranked* the translations for 60 sentences. We also focused on the outputs of the two *de*→*hsb* and *de*→*dsb* translation directions. Given the poor performance of the baseline, we only consider this year’s four submitted systems (HeyBusan, HITSZ, HSE, and LT3) and last year’s overall winning system, TartuNLP, for a total of five system translations per sentence. Ties were allowed in the ranking process. Further details on the annotation methodology are presented in last year’s annotation explanation (Okabe et al., 2025).

Table 14 summarises the ranking of the five systems between ranks obtained through chrF++ scores and assigned by human annotators. We observe that human evaluation agrees with the automatic evaluation overall, with the TartuNLP system having the best translation among the five. This year’s submissions then follow less distinctly. It is worth noting that for Lower Sorbian, MT outputs still have room for improvement, as the best translation among the five systems is sometimes ranked 2 rather than 1.

6.3 Detailed performance for Sorbian QA

Table 15 displays the QA accuracy score per language proficiency level in both Sorbian languages. As the test set is the same as in the 2025 edition, we additionally report the scores of the previous four participating systems: TartuNLP, SDKM (Saadi et al., 2025), the baseline Qwen2.5-3B, and NRC (Larkin et al., 2025). First, we note that this year’s Qwen3.5-2B base model is comparable to last year’s, with only a slight loss in accuracy.

Then, we observe that this year’s best system for QA (LT3) outperforms last year’s TartuNLP, setting a new state of the art for the task. The main improvement for LT3 comes from the two intermediate B levels (B1 and B2). For both Sorbian languages, it outperforms TartuNLP by at least 15 points of accuracy. Starting at the B1 level, the language certification features an additional exercise category: grammatical questions (on top of the listening and reading comprehension). Although Moerman and Tezcan (2026) have a dedicated training phase to improve grammar-checking performance, they report similar QA scores before and after it.

	rank	chrF++					human				
		Tartu	HeyB.	LT3	HITSZ	HSE	Tartu	HeyB.	LT3	HITSZ	HSE
Upper Sorbian	1st	48	8	3	1	0	45	12	3	0	0
	2nd	7	24	13	10	6	10	21	8	9	7
	3rd	2	21	19	7	11	3	21	24	11	5
	4th	3	3	16	23	15	1	5	20	22	13
	5th	0	4	9	19	28	1	1	5	18	35
Lower Sorbian	1st	41	9	5	0	5	27	8	2	3	0
	2nd	6	20	15	7	12	21	19	14	1	4
	3rd	4	23	15	8	10	7	20	28	10	9
	4th	6	8	15	18	13	5	8	10	21	23
	5th	3	0	10	27	20	0	5	6	25	24

Table 14: System rankings according to the chrF++ score and the human evaluation for 60 sentences. Tartu stands for the TartuNLP system, while HeyB. is for the HeyBusan team. For instance, the TartuNLP translations in Upper Sorbian were ranked first among the five systems for 48 sentences according to chrF++, while it was 45 times according to the human evaluation.

model	A1	A2	B1	B2	C1	all.
<i>Upper Sorbian</i>						
TartuNLP	86.67	82.14	56.82	37.50	51.92	58.10
SDKM	80.00	78.57	56.82	41.07	42.31	55.24
Qwen2.5-3B	70.00	57.14	40.91	26.79	38.46	42.86
NRC	50.00	32.14	22.73	19.64	30.77	29.05
LT3	83.33	75.00	81.82	64.29	40.38	66.19
HITSZ	96.67	71.43	47.73	48.21	50.00	58.57
HeyBusan	93.33	78.57	43.18	41.07	44.23	54.76
HSE	63.33	64.29	45.45	30.36	38.46	44.76
baseline	60.00	64.29	40.91	26.79	38.46	42.38
<i>Lower Sorbian</i>						
TartuNLP	89.66	71.43	56.82	41.07	50.00	57.56
SDKM	82.76	57.14	50.00	37.50	47.92	51.71
Qwen2.5-3B	65.52	75.00	43.18	26.79	41.67	45.85
NRC	55.17	42.86	18.18	26.79	31.25	32.20
LT3	93.10	67.86	79.55	57.14	37.50	63.90
HITSZ	93.10	67.86	47.73	46.43	45.83	56.10
HeyBusan	93.10	75.00	45.45	46.43	47.92	57.07
HSE	65.52	57.14	47.73	37.50	47.92	48.78
baseline	65.52	50.00	43.18	33.93	41.67	44.39

Table 15: Accuracy on the QA datasets per language level for **Upper** and **Lower Sorbian**. The upper part of a table for a language lists the systems from the 2025 edition, while the lower part shows this year’s shared task systems. Models are ordered based on their overall QA task score per year. The best score per edition and level is in **bold**, while the overall best is underlined.

We hence suppose that LT3’s first multitask training step, combined with their inference strategy, was already helpful to reach such QA scores.

We also note that the other systems show a smaller but notable improvement for the B level compared to the baseline. Further understanding of the synergy between the GC task and the QA score

improvement is left for future analyses. Finally, if we keep a passing accuracy of 50 for a language certification, as in last year’s edition, most systems sit somewhere in the intermediate A2–B1 level, with LT3 being at B2.²¹

6.4 Detailed analysis for spell checking

Table 16 presents the SC detection and correction accuracy values for all participating systems and the three languages per modification type. The baseline zero-shot model performs poorly on the task. The outputs tend not to follow the formatting instructions (as seen in the differing scores between detection and correction for correct sentences), on top of failing to produce correct answers. All the submitted systems performed better due to their task-specific training. We then note that error detection remains easier than error correction overall, with notably higher accuracy values. Finally, the less naive transformations remain challenging with accuracy values below 70%, compared to the other operations. In our case, deletion focused mainly on the later part of a word (e.g., a truncation of a word), while subtle substitutions led to similar but wrong words.

6.5 Detailed analysis for grammar checking

Table 17 presents the GC detection and correction accuracy for all systems and languages, looking at errors introduced in nouns and adjectives separately. The baseline model performs poorly across all languages, as do the systems for Ukrainian. Notably,

²¹Actual language certifications feature more exercise types.

team	language	correct		deletion		insertion		random subst.		subtle subst.		swapping	
		D	C	D	C	D	C	D	C	D	C	D	C
Ukrainian track													
Zolint	uk	86.2	86.2	17.5	10.0	77.0	51.0	74.5	33.5	43.0	33.5	74.0	52.5
koshi	uk	58.2	58.2	32.0	10.0	77.0	40.0	78.0	28.5	50.0	35.0	73.0	44.5
baseline	uk	14.4	14.8	1.0	1.0	3.5	2.0	3.0	2.5	1.5	1.0	4.5	1.5
TUM-HN	uk	0.0	0.0	21.5	9.5	55.0	26.5	52.5	20.0	34.5	22.5	58.5	29.5
Sorbian track													
HeyBusan	hsb	89.5	89.5	53.0	35.5	92.0	65.5	91.5	53.0	66.5	55.0	92.0	78.5
	dsb	89.6	89.6	53.0	32.5	89.5	65.0	89.0	53.0	55.0	47.5	89.0	74.0
LT3	hsb	89.9	89.9	59.0	34.0	83.0	53.5	85.0	48.5	61.5	51.5	81.0	59.5
	dsb	86.8	86.8	60.0	39.0	83.5	57.5	87.0	58.5	64.5	55.0	86.5	69.5
HITSZ	hsb	90.8	90.8	45.5	29.0	86.5	50.5	84.5	42.5	58.5	35.0	86.5	58.5
	dsb	92.4	92.4	36.0	26.5	82.0	54.0	79.0	39.0	43.5	28.0	78.0	50.0
HSE	hsb	80.6	80.6	27.5	6.5	81.5	39.5	84.5	17.0	49.5	23.5	78.0	39.0
	dsb	82.2	82.2	31.5	14.0	84.0	44.0	81.0	25.5	43.0	28.0	76.5	43.0
baseline	hsb	12.7	13.4	1.0	0.0	1.5	0.0	1.0	0.0	1.0	0.5	1.5	0.5
	dsb	11.9	12.6	1.5	0.0	1.0	0.0	2.5	0.0	0.5	0.0	1.0	0.0

Table 16: **Spelling** error **detection** (D) and **correction** (C) scores per team and transformation type. Cells are shaded according to the accuracy values (the darker, the better). Systems are ordered by their final ranking.

Zolint nearly always predicts no error, whereas TUM-HN always predicts an error. We attribute the poor performance on the Ukrainian task to the dataset, which contains large portions of formal and legal language, making this test set comparatively difficult.

For Sorbian, we observe better performance, potentially helped by the domain similarity between the test set and the MT training data. As for spell checking, error detection is easier than generating a correction. Between adjectives and nouns, we do not observe a clear difference, even though the HITSZ results are considerably worse for nouns than for adjectives. In contrast, both HeyBusan and LT3 show higher scores in detecting and correcting errors in nouns.

Overall, the task of grammar checking (both error detection and correction) remains challenging.

6.6 Inspecting MR responses

We inference the submitted models on the respective MR tasks with the evaluation harness and inspect their outputs. To get a general sense of the outputs, we manually inspect the first 15 responses

per model and language and observe some patterns. As in the zero-shot baseline, the majority of submitted models attempt to reason through the maths problems even when ‘thinking’ mode is turned off. The exceptions are the models from HITSZ and HSE, which only output the answer format in this setting. In addition, as in the zero-shot baseline, the models from TUM-HN and Zolint frequently overran the output token budget of 512 before committing to an answer, but when inferencing the TUM-HN model with a token budget of 2048, it gave a correctly formatted answer in most of the inspected responses. The models from HeyBusan and LT3 answered within the shorter token budget much more consistently.

We further observed that the HITSZ model still only outputs the answer format when ‘thinking’ mode is turned on, while the HSE model displays mixed behaviour, sometimes answering directly, sometimes generating reasoning traces. The HSE and the Zolint models also sometimes entered decoding loops after answering, which, for the HSE model, seemed to become more frequent when the thinking mode was turned on. Since everything

team	lang.	correct		nouns		adjectives	
		D	C	D	C	D	C
Ukrainian track							
Zolint	uk	90.2	90.2	0.3	0.0	0.3	0.0
koshi	uk	43.5	43.5	31.4	4.6	25.3	8.0
baseline	uk	1.2	2.9	4.3	0.4	4.0	0.3
TUM-HN	uk	0.0	0.0	24.7	6.9	11.3	5.7
Sorbian track							
HeyBusan	hsb	90.9	90.9	73.2	59.5	58.9	52.5
	dsb	92.6	92.6	79.2	68.1	70.5	63.3
LT3	hsb	91.7	91.7	72.4	59.2	57.6	48.1
	dsb	90.4	90.4	78.6	68.3	63.3	51.8
HITSZ	hsb	94.5	94.5	22.2	17.1	55.0	42.9
	dsb	94.9	94.9	22.5	16.1	68.1	61.4
HSE	hsb	99.9	99.9	0.5	0.0	0.8	0.3
	dsb	99.8	99.8	0.1	0.0	0.0	0.0
baseline	hsb	0.3	0.6	3.6	0.2	3.4	0.0
	dsb	0.3	0.9	5.7	0.0	1.8	0.0

Table 17: **Grammar error detection (D) and correction (C) scores per team and part-of-speech tag.** Cells are shaded according to the accuracy values (the darker, the better). Systems are ordered by their final ranking. Note that adjectives and nouns are not equally distributed. UKR: 700 nouns, 300 adjectives; HSB: 613 nouns, 387 adjectives; DSB: 697 nouns, 166 adjectives.

but the response within the answer format tags was stripped out before evaluation, such decoding loops, or which language the model used for ‘reasoning’, did not have a direct effect on performance.

Language identification However, it is possible that the reasoning language has an indirect effect on model performance. Therefore, we run language identification with GlotLID (Kargaran et al., 2023) over the model outputs where reasoning traces were present. In the Ukrainian track, the models typically output Ukrainian (84% from the baseline, 75% and 78% from Zolint and TUM-HN), with a smaller proportion of outputs classified as English (baseline 15%, TUM-HN 21%, none from Zolint), Russian (2% from Zolint), or a wide selection of other languages. We assume that these other languages are most often spurious classifications for outputs consisting mostly of symbols.

In the Sorbian track, when models output reasoning tokens, they are most commonly in English. The baseline used English reasoning traces around

47% of the time, while HeyBusan and LT3 did so between 82% and 86% of the time, with only small variation between Upper and Lower Sorbian inputs. The baseline’s reasoning traces given Upper Sorbian input were also classified as Upper Sorbian 25% of the time, with the remainder split between Polish (12%), Czech (8%), and a few other Slavic languages. For Lower Sorbian inputs, the classifier still found 14% Upper Sorbian traces, compared to 10% Lower Sorbian, and 20% Polish. This variation largely disappears in the LT3 and HeyBusan models, consistent with the described training sets. In combination with the consistent, faster answers, this likely explains the high performance of these models on Sorbian MR.

7 Conclusion

This edition of the Shared Task extended the number of tasks to five, compared to the previous edition, by covering: Machine Translation, Question Answering, Spell Checking, Grammar Checking, and Maths Reasoning. The base model was also updated to Qwen3.5-2B, a more recent version but with fewer parameters (compared to Qwen2.5-3B).

First, we observe that the participating systems outperformed the baseline for most tasks overall, sometimes by a clear margin for the better systems. This is naturally due to task-specific training, especially for tasks with specific formatting (SC and GC). A single system to handle the five tasks for Ukrainian, Upper and Lower Sorbian is hence possible.

For Sorbian MT, we observed generally higher quality than for Ukrainian MT, as Sorbian training and test data are from similar domains, while the Ukrainian test set contains paragraphs which proved to be a hurdle for the models. Moreover, for Sorbian, the ranking with xCOMET-XXL agrees with the surface-level metric (chrF++), whereas for Ukrainian, the first-ranked system’s lead was considerably reduced under xCOMET. Furthermore, the small-scale manual evaluation for the two Sorbian languages confirmed the human agreement with chrF++ for translation rankings. A comparison of translation quality for Sorbian with last year’s winning team (TartuNLP) showed that this year’s performance remains below last year’s best performance. However, TartuNLP’s system is based on a slightly larger model (3B) and trained on fewer tasks, making this comparison not entirely fair.

For QA, all systems outperformed the baseline. This year’s best systems showed improved performance compared to last year’s winners on the same QA test sets. Regarding the three new tasks, the baseline model exhibited output formatting issues for both spell and grammar checking and general difficulty for maths reasoning in a zero-shot setting. Despite its simple setting, spell checking is not entirely solved by the submitted systems. While systems can detect errors in a sentence, correcting remains more difficult, especially for subtle errors in all three languages. Similarly, grammar checking remains challenging, with correction being systematically harder than detection. Finally, maths reasoning proved to be the hardest task for the systems, with the best scores in the 20–30 accuracy range for low- to medium-level questions.

In short, we overall observed better task performance than the baseline with the submitted systems for both tracks. Most of them are also capable of performing the five tasks, with minimal sacrifice in performance on tasks that are not necessarily related. A cautious multitask training approach seems, however, necessary, as catastrophic forgetting was reported otherwise. The submissions from the participants did not show clear evidence of synergy between the tasks themselves. Since the training mixture was usually heavily imbalanced towards MT compared to the other tasks due to data scarcity, especially for the Sorbian track, no general trend can be drawn. Yet, for a given training data and strategy in two systems, focusing on MT has been shown to negatively affect QA, as in last year. Finally, on the overlapping MT and QA tasks, this year’s best submission for Ukrainian outperforms last year’s, while for Sorbian, TartuNLP remains the best open-source system for MT.

This shared task showed that the three new tasks remain challenging, especially for MR. Several factors were explored by the participants, which need further investigation in terms of their final impact: the quality and distribution of each task in the dataset, task-specific inference strategies, or the use of reasoning capabilities. This year’s edition displayed promising results for small but multitask LLMs in underrepresented languages.

Ethics Statement

The shared task continued to focus on smaller models: this year’s base model, Qwen3.5-2B, is still below the milestone of 3B parameters from last

year. This ensures that participants have a lower computational barrier, fostering both accessibility and a smaller ecological footprint. We notably have two participants using consumer-grade GPUs for training and inference.

We also commit to open science and reproducibility by ensuring that the teams use both open-source models and publicly available datasets. Participants were encouraged to make their model public, and all systems from this year are on HuggingFace.

For the Sorbian track, the shared task relies on the partnership with the WITAJ-Sprachzentrum. They provided all new Sorbian datasets for MT and annotation help for the PluraMath dataset for MR. For the manual evaluation experiment on the Sorbian MT outputs, both annotators were contacted through the WITAJ-Sprachzentrum.

Limitations

The main limitation comes from the restriction to one model size and family. Exploring a wider range of model families in that size range might help establish the extent to which the presented strategies are generalisable across models, in particular regarding issues observed with the base Qwen3.5-2B model, such as decoding loops in the thinking mode. For this edition, we constrained the model to a single base model for a fair comparison between participating teams.

The Shared Task continues to focus on three languages (Ukrainian, Upper Sorbian, and Lower Sorbian) with two dedicated language tracks. The methodological findings from the submitted systems are hence only applicable to these languages within the shared task setting (including the five tasks and their specific formatting).

Acknowledgments

This work was partly funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the

NHR project b279bb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

The authors also gratefully acknowledge the scientific support and resources of the AI service infrastructure *LRZ AI Systems* provided by the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW), funded by Bayerisches Staatsministerium für Wissenschaft und Kunst (StMWK).

We thank the UNLP 2024 Shared Task team: Roman Kyslyi, Mariana Romanyshyn, and Oleksiy Syvokon, for sharing the Ukrainian QA resources. We also thank Hanna Yukhymenko for the support with the MMLU_ukr QA dataset.

Moreover, we are grateful for our cooperation with the WITAJ-Sprachzentrum for Upper and Lower Sorbian. In the continuation of the datasets they provided for the previous WMT shared tasks on MT for low-resource languages and last year’s QA data, they provided new datasets for this year’s edition. They have helped us with the PluraMath project for both Sorbian MR splits that we use in this Shared Task. We thank Tomaš Šolta for providing the language certificate exercises and Kryštof Peršín for the maths reasoning dataset annotation. We also thank the annotators of Lower Sorbian MT outputs for their time and comments.

We warmly thank the soblexx team, in particular Bernhard Baier, for his trust and help regarding the Upper Sorbian linguistic resource to create our grammar checking dataset.

Finally, we thank the organisers of the WMT 2026 shared tasks. We highly appreciate the help from Martin Popel, Tom Kocmi, Katia Artemova, and Mariya Shmatova for sharing the test sets for the Ukrainian MT track with us in 2025. We are also especially grateful to Philipp Koehn and Christof Monz for helping us with Softconf, and Roman Grundkiewicz for setting up again the leaderboard on OCELoT for our shared task.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual evaluation of generative AI](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

4232–4267, Singapore. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Daryna Dementieva, Nikolay Babakov, Kathy Hämerl, Ilseyar Alimova, Jindřich Libovický, Shu Okabe, Miras Baisbay, Lukas Edman, Abrorkhon Inomkhujayev, Antonia Karamolegkou, Mateusz Lango, Volkan Özer, Nikola Selic, Subhankar Swain, Tsedeniya Kinfe Temesgen, Galit Bary Weisberg, and Alexander Fraser. 2026. [PluraMath: Extending mathematical reasoning evaluation beyond high-resource languages](#). *Preprint*, arXiv:2607.05992.

Daryna Dementieva, Valeriia Khylenko, and Georg Groh. 2025. [Cross-lingual text classification transfer: The case of Ukrainian](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1451–1464, Abu Dhabi, UAE. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

Michael Diskin. 2026. One Model, Five Tasks, Two RTX 3090 GPUs: The HSE System for the Sorbian Track of WMT26 Multitask LLMs with Limited Resources. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Nazarii Drushchak and Mariana Romanyshyn. 2024. [Introducing the Djinni recruitment dataset: A corpus of anonymized CVs and job postings](#). In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 8–13, Torino, Italia. ELRA and ICCL.

Mahbub E Sobhani, Md. Faiyaz Abdullah Sayeedi, Tasnim Mohiuddin, Md Mofijul Islam, and Swakkhar

- Shatabda. 2026. [MathMist: A parallel multilingual benchmark dataset for mathematical problem solving and reasoning](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2524–2550, Rabat, Morocco. Association for Computational Linguistics.
- 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.
- Marco Gaido, Roman Grundkiewicz, Thamme Gowda, and Matteo Negri. 2025. [Findings of the WMT 2025 shared task on model compression: Early insights on compressing LLMs for machine translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 484–494, Suzhou, China. 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.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the math dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. [ua_datasets: A collection of Ukrainian language datasets](#).
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Mikhail Khasikov. 2026. [HeyBusan at WMT26: A System for the Sorbian Track](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025a. [Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025b. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- 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*, pages 1133–1142, 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.
- Liquid AI. 2026. [Lfm2.5-1.2b-thinking: On-device reasoning under 1gb](#). *Liquid AI Blog*. www.liquid.ai/blog/lfm2-5-1-2b-thinking-on-device-reasoning-under-1gb.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien

- Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. [Ministral 3](#). *Preprint*, arXiv:2601.08584.
- Thomas Moerman and Arda Tezcan. 2026. Getting More from Small LLMs: Fuzzy-Match Retrieval and Candidate Scoring for Multitask Sorbian NLP. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Shu Okabe, Daryna Dementieva, Marion Di Marco, Lukas Edman, Katharina Haemmerl, Marko Měškank, Anita Hendrichowa, and Alexander Fraser. 2025. [Findings of the WMT 2025 shared task LLMs with limited resources for Slavic languages: MT and QA](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 503–519, Suzhou, China. Association for Computational Linguistics.
- Partha Pakray, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. [Findings of WMT 2025 shared task on low-resource Indic languages translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. 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.
- 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*, pages 1143–1150, Suzhou, China. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- 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, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Nathaniel Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Kenton Murray, Raj Dabre, Andre Coy, and Heather Lent. 2025. [Findings of the first shared task for creole language machine translation at WMT25](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 520–531, Suzhou, China. Association for Computational Linguistics.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. 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*, pages 1151–1157, Suzhou, China. Association for Computational Linguistics.
- Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. [UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language](#). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Tsedeniya Kinfe Temesgen, Sinan Kurtyigit, and Alexander Fraser. 2026. [TUMHN@WMT2026: LLMs with Limited Resources for Slavic Languages](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multi-lingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

- Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Polymath: Evaluating mathematical reasoning in multilingual contexts](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- 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.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Lingchu Yang, Zitong Fu, Kathy Hämmerl, and Masaki Ito. 2026. Zolint at WMT 2026: A Multitask LLM Submission for the Low-Resource Ukrainian Track. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. [MamayLM: An efficient state-of-the-art Ukrainian LLM](#).
- Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2026. [Recovered in translation: Efficient pipeline for automated translation of benchmarks and datasets](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41590–41609, San Diego, California, United States. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.
- Cong Zhang, Yutong Wang, Xuebo Liu, and Min Zhang. 2026. HITSZ at WMT 2026: Mixed Continued Pre-training and Supervised Fine-tuning for Low-Resource Sorbian. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025. [Hy-mt1.5 technical report](#). *Preprint*, arXiv:2512.24092.

A Prompts for the zero-shot baseline

```
Translate the following German text to Lower
↳ Sorbian. Put it in this format <dsb>
↳ Lower Sorbian translation </dsb>.
<deu> {{source}} </deu>
```

Figure 4: Baseline MT prompt, here for de→dsb.

```
{{context}}

Question:
{{question}}

Possible answers:
{{possible_answers}}

Answer:
```

Figure 5: Baseline question answering prompt.

```
The following sentence contains at most one
↳ misspelled word. Respond with exactly
↳ these two tags and nothing else:
<wrong> misspelled word </wrong> <corrected>
↳ correct spelling </corrected>
If the sentence has no error, use the literal
↳ string CORRECT in both tags:
<wrong> CORRECT </wrong> <corrected> CORRECT
↳ </corrected>
<sentence> {{input_sentence}} </sentence>
```

Figure 6: Baseline prompt for spell checking.

```
The following sentence contains at most one
↳ grammatically incorrect word. Respond
↳ with exactly these two tags and
↳ nothing else:
<wrong> misspelled word </wrong> <corrected>
↳ correct spelling </corrected>
If the sentence has no error, use the literal
↳ string CORRECT in both tags:
<wrong> CORRECT </wrong> <corrected> CORRECT
↳ </corrected>
<sentence> {{input_sentence}} </sentence>
```

Figure 7: Baseline prompt for grammar checking.

B Detailed Grammar Checking example

Figure 9 shows a Ukrainian example sentence for the grammar check, where the grammatical case

```
Answer this mathematical reasoning question.
↳ Place your final answer, using LaTeX
↳ when appropriate, in this format:
↳ <answer> answer </answer> {{question}}
```

Figure 8: Baseline prompt for mathematical reasoning.

(*INS*) of the noun *результатом* ('results') has been changed into *GEN*, leading to a corrupted sentence.

C Detailed results tables

Tables 18 and 19 display the detailed results obtained by the participating systems per sub-task for the Ukrainian and Sorbian tracks, respectively.

D Statistical test of the final ranking

Our official results only use the output scores to compute each task ranking of the systems. As some of the differences are very minor, we perform statistical tests on each of the five tasks to assess the stability of our ranking. Several tasks feature sub-tasks (e.g., MT, QA, and the remaining Sorbian tasks), which hence require a stratified approach. We note that for Sorbian, most questions in QA and all the MR exercises are translations of each other; they are hence not independent and need to be considered together. We report the results of a stratified paired cluster bootstrap with 20,000 replicates for the confidence interval. We additionally provide the rank range for each task based on the significance of the difference. We use a paired approximate randomisation test between the systems at 5% with Holm correction. We compute the final confidence interval and the rank range using the same official point methodology on each bootstrap resample per task.

Tables 20 and 21 show the submission ranks after our resampling analysis. When systems cannot be separated statistically for a task, we provide a ranking range (e.g., 1–2 for first and second positions). We do not observe a major difference in the official final rank for both tracks: the tied winning systems for the Sorbian track are indeed not significantly different, while for the Ukrainian track, the baseline and the TUM-HN systems are also indistinguishable. We note here that Sorbian QA is the task where separating systems is the most difficult in our case.

original sentence	Наукові факти є результатом експериментів.
corrupted sentence	Наукові факти є результату експериментів.
target lemma	результат
target word + features	результатом/N;INS;SG;INAN;MASC
swapped word + features	результату/N;GEN;SG;INAN;MASC
swapped features	NOUN; INS → GEN
gloss	Scientific facts are the result of experiments.

Figure 9: Grammar checking: example sentence for Ukrainian.

Task	Subset	Metric	TUM-HN	baseline	koshi	Zolint
MT	all	chrF++	16.08	18.46	16.80	49.86
	ces-ukr	BLEU	3.93	5.38	6.01	23.66
	ces-ukr	chrF++	19.77	25.44	23.05	51.53
	eng-ukr	BLEU	1.01	0.55	0.70	15.43
	eng-ukr	chrF++	12.39	11.47	10.55	48.18
QA	all	Exact Match	40.13	36.33	46.99	41.15
	ukr-ZNO	Exact Match	35.82	30.76	45.01	37.42
	ukr-MMLU	Exact Match	44.44	41.90	48.97	44.88
SC	all	Exact Match	16.50	8.38	52.50	66.43
	ukr-SC	Exact Match (Wrong)	22.20	8.55	60.10	71.70
		Exact Match (Corrected)	10.80	8.20	44.90	61.15
GC	all	Exact Match	6.80	2.18	30.55	45.18
	ukr-GC	Exact Match (Wrong)	10.35	2.70	36.55	45.25
		Exact Match (Corrected)	3.25	1.65	24.55	45.10
MR	ukr-MR	Exact Match	4.40	20.00	0.40	21.60

Table 18: Results per sub-task for the **Ukrainian** track. Systems are ordered based on the main ranking from left to right. ‘all’ indicates the averaged task score.

Task	Subset	Metric	baseline	HSE	HITSZ	LT3	HeyBusan
MT	all	chrF++	21.68	61.50	63.12	69.57	72.75
	deu-hsb	BLEU	0.96	30.48	35.32	42.07	48.61
	deu-hsb	chrF++	13.67	56.27	60.04	64.55	70.26
	hsb-deu	BLEU	6.08	27.14	37.42	44.11	46.62
	hsb-deu	chrF++	27.05	52.70	62.42	67.19	69.43
	deu-dsb	BLEU	0.95	27.14	26.87	34.56	39.36
	deu-dsb	chrF++	10.42	52.72	52.37	58.90	63.06
	dsb-deu	BLEU	3.73	29.90	36.30	44.05	47.87
	dsb-deu	chrF++	23.18	54.61	60.92	66.56	69.42
	dsb-hsb	BLEU	5.81	55.78	49.42	63.18	66.67
	dsb-hsb	chrF++	28.66	75.50	70.62	80.09	82.38
	hsb-dsb	BLEU	4.88	57.32	50.50	62.12	64.82
	hsb-dsb	chrF++	27.10	77.19	72.35	80.12	81.96
QA	all	Exact Match	43.39	46.77	57.33	65.05	55.92
	hsb-QA	Exact Match	42.38	44.76	58.57	66.19	54.76
	dsb-QA	Exact Match	44.39	48.78	56.10	63.90	57.07
SC	all	Exact Match	6.66	63.63	73.13	76.11	78.03
	hsb-SC	Exact Match (Wrong)	6.95	72.40	81.55	81.90	84.25
		Exact Match (Corrected)	6.80	52.85	66.95	69.65	73.50
	dsb-SC	Exact Match (Wrong)	6.60	72.70	78.05	81.55	82.35
		Exact Match (Corrected)	6.30	56.55	65.95	71.35	72.00
GC	all	Exact Match	1.34	50.02	62.11	78.32	79.52
	hsb-GC	Exact Match (Wrong)	1.90	50.25	64.70	79.20	79.30
		Exact Match (Corrected)	0.35	50.00	60.80	73.30	73.85
	dsb-GC	Exact Match (Wrong)	2.67	49.94	63.09	83.02	85.05
		Exact Match (Corrected)	0.46	49.88	59.85	77.75	79.90
MR	all	Exact Match	5.40	5.00	18.60	30.40	29.00
	hsb-MR	Exact Match	6.00	5.20	19.20	32.00	29.20
	dsb-MR	Exact Match	4.80	4.80	18.00	28.80	28.80

Table 19: Results per sub-task for the **Sorbian** track. Systems are ordered based on the main ranking from left to right. ‘all’ indicates the averaged task score.

task	system	score	95% CI	rank
MT	Zolint	49.86	[48.04, 51.16]	1
	baseline	18.46	[16.60, 21.23]	2
	koshi	16.80	[14.84, 19.65]	3–4
	TUM-HN	16.08	[13.80, 19.51]	3–4
QA	koshi	46.99	[45.16, 48.82]	1
	Zolint	41.15	[39.39, 42.93]	2–3
	TUM-HN	40.13	[38.38, 41.90]	2–3
	baseline	36.33	[34.62, 38.04]	4
SC	Zolint	66.43	[64.55, 68.30]	1
	koshi	52.50	[50.52, 54.48]	2
	TUM-HN	16.50	[15.10, 17.95]	3
	baseline	8.38	[7.23, 9.57]	4
GC	Zolint	45.18	[43.05, 47.35]	1
	koshi	30.55	[28.70, 32.43]	2
	TUM-HN	6.80	[5.88, 7.78]	3
	baseline	2.18	[1.68, 2.73]	4
MR	Zolint	21.60	[16.40, 26.80]	1–2
	baseline	20.00	[15.20, 25.20]	1–2
	TUM-HN	4.40	[2.00, 7.20]	3
	koshi	0.40	[0.00, 1.20]	4
final	Zolint	19	[17, 19]	1
	koshi	13	[12, 13]	2
	baseline	9	[9, 10]	3–4
	TUM-HN	9	[9, 11]	3–4

Table 20: System ranking with significance clusters for the **Ukrainian track**. We report the official score, the confidence interval (CI) at 95%, and the resulting rank.

task	system	score	95% CI	rank
MT	HeyBusan	72.75	[72.56, 72.94]	1
	LT3	69.57	[69.37, 69.76]	2
	HITSZ	63.12	[62.92, 63.32]	3
	HSE	61.50	[61.30, 61.70]	4
	baseline	21.68	[21.51, 21.85]	5
QA	LT3	65.05	[59.52, 70.42]	1–3
	HITSZ	57.33	[51.19, 63.38]	1–3
	HeyBusan	55.92	[49.80, 61.89]	1–4
	HSE	46.77	[40.53, 52.87]	3–5
	baseline	43.39	[37.50, 49.26]	4–5
SC	HeyBusan	78.03	[76.83, 79.20]	1
	LT3	76.11	[74.88, 77.32]	2
	HITSZ	73.13	[71.86, 74.36]	3
	HSE	63.63	[62.28, 64.95]	4
	baseline	6.66	[5.92, 7.43]	5
GC	HeyBusan	79.52	[78.28, 80.76]	1–2
	LT3	78.32	[77.03, 79.56]	1–2
	HITSZ	62.11	[60.56, 63.63]	3
	HSE	50.02	[48.43, 51.64]	4
	baseline	1.34	[1.08, 1.63]	5
MR	LT3	30.40	[25.20, 35.80]	1–2
	HeyBusan	29.00	[23.80, 34.40]	1–2
	HITSZ	18.60	[14.40, 23.00]	3
	baseline	5.40	[3.40, 7.80]	4–5
	HSE	5.00	[2.80, 7.40]	4–5
final	HeyBusan	22	[22, 24]	1–2
	LT3	22	[21, 23]	1–2
	HITSZ	16	[15, 16]	3
	HSE	9	[8, 10]	4
	baseline	6	[5, 7]	5

Table 21: System ranking with significance clusters for the **Sorbian track**. We report the official score, the confidence interval (CI) at 95%, and the resulting rank.