

HITSZ at WMT 2026: Mixed Continued Pre-training and Supervised Fine-tuning for Low-Resource Sorbian

Cong Zhang, Yutong Wang, Xuebo Liu*, Min Zhang

Harbin Institute of Technology, Shenzhen, China

{zhangcong, wangyutong}@stu.hit.edu.cn

{liuxuebo, zhangmin2021}@hit.edu.cn

Abstract

We describe the HITSZ system submitted to the Sorbian track of the WMT 2026 Shared Task on Multitask LLMs with Limited Resources. The track requires a single Qwen3.5-2B model to jointly perform machine translation (MT), question answering (QA), spell checking (SC), grammar checking (GC), and mathematical reasoning (MR) for Upper and Lower Sorbian. To strengthen the model’s limited Sorbian linguistic knowledge while retaining previously acquired capabilities, we first perform mixed continued pre-training (CPT) on Sorbian monolingual and parallel data together with multilingual and task-oriented replay. We then apply full-parameter multitask supervised fine-tuning (SFT) to align the adapted model with the five heterogeneous task formats. Compared with the official Qwen3.5-2B baseline, our primary submission improves MT by 41.44 chrF++ points and improves QA, SC, GC, and MR by 13.94, 66.47, 60.77, and 13.20 points, respectively, ranking third overall in the Sorbian track. Additional analysis shows that removing CPT reduces our five-task development score by 5.24 points, with the largest losses on SC, MT, and GC. We also find that heavily increasing MT supervision does not further improve translation and can degrade other jointly trained tasks.

1 Introduction

Large language models (LLMs) have substantially advanced multilingual natural language processing, but their capabilities remain uneven across languages, particularly when high-quality text and task supervision are scarce. Upper Sorbian (hsb) and Lower Sorbian (dsb) are two closely related West Slavic minority languages spoken primarily in eastern Germany, with approximately 30,000 and 7,000 native speakers, respectively (Okabe et al., 2025). Their limited digital resources make

it difficult for compact multilingual models to acquire reliable linguistic and downstream-task capabilities. The WMT 2025 shared task studied this problem through joint machine translation (MT) and question answering (QA) (Okabe et al., 2025). WMT 2026 broadens the Sorbian setting to five task families—MT, QA, spell checking (SC), grammar checking (GC), and mathematical reasoning (MR)—which must be handled by a single Qwen3.5-2B model (Okabe et al., 2026; Qwen Team, 2026).

The WMT 2025 systems demonstrate both the potential and the difficulty of low-resource Sorbian adaptation. TartuNLP (Purason and Fishel, 2025) continued training Qwen2.5-3B-Instruct on a mixture of Sorbian monolingual, parallel, and general instruction data, obtaining the highest overall ranking across MT and QA. NRC (Larkin et al., 2025) instead focused primarily on MT with full-parameter supervised fine-tuning and achieved the strongest translation results, but substantially lower QA accuracy, while JGU Mainz (Saadi et al., 2025) jointly adapted MT and QA using parameter-efficient fine-tuning and additional task data. The shared-task findings consequently observed that training purely for MT can degrade the model’s original QA capability (Okabe et al., 2025). These studies addressed the two-task MT–QA setting; the addition of SC, GC, and MR in WMT 2026 creates a broader requirement in which language-specific knowledge and general task capabilities must coexist within the same compact model.

We address this requirement with a two-stage adaptation strategy. First, the base model has limited Sorbian linguistic competence, while continued pre-training on a narrow target-language corpus risks catastrophic forgetting of capabilities acquired during pre-training. Replay of general pre-training data has been used in recent MT-oriented continued pre-training to mitigate this effect (Zheng et al., 2025). Motivated by this,

*Corresponding author.

our *mixed CPT* combines Sorbian monolingual and parallel data with English and German replay, directional MT supervision, instruction data, mathematical reasoning examples, and lightweight downstream-task anchors. Second, the five downstream tasks differ substantially in output format, supervision scale, and optimization behavior. Starting from the selected CPT checkpoint, we therefore perform full-parameter multitask SFT with assistant-only loss and explicitly control task exposure rather than allowing the largest training source to dominate the supervised-token budget. Together, the two stages separate the acquisition of stronger Sorbian language competence from the subsequent alignment of that competence with heterogeneous task requirements.

Our primary submission ranks third overall in the WMT 2026 Sorbian track, achieving 63.12 chrF++ on MT, 57.33 on QA, 73.13 on SC, 62.11 on GC, and 18.60 on MR (Okabe et al., 2026). The final submitted checkpoint is publicly available on HuggingFace.¹ Compared with the official Qwen3.5-2B baseline, the system improves all five task scores, with particularly large gains on SC, GC, and MT. Our development experiments further provide two observations. Removing mixed CPT while keeping the downstream SFT recipe unchanged reduces the five-task development score from 57.45 to 52.21, with the largest losses on SC, MT, and GC. In addition, increasing MT to nearly three quarters of the supervised-token budget does not improve translation and reduces performance on several other tasks. These results suggest that effective low-resource multitask adaptation requires both target-language adaptation and careful allocation of supervision across downstream capabilities.

2 Data

2.1 Mixed Continued Pre-training Data

The base Qwen3.5-2B model has limited exposure to Sorbian because the language is severely under-represented in large-scale pre-training corpora. We therefore first perform continued pre-training to increase Sorbian linguistic competence before downstream task alignment. However, directly optimizing on a narrow low-resource corpus may cause over-specialization and degradation of previously acquired multilingual capabilities. To address this

Component	Tokens	Role
Sorbian monolingual	110.87M	adaptation
Parallel bilingual text	4.22M	alignment
Directional MT replay	18.60M	MT replay
English monolingual	13.53M	replay
German monolingual	9.29M	replay
Sorbian instructions	7.62M	instruction
Mathematical CoT	6.49M	reasoning
Task anchors	16.12M	task replay
Total	186.74M	

Table 1: Composition of the mixed CPT corpus. Token counts denote full sequence tokens; assistant-only components contain fewer trainable tokens.

trade-off, our CPT data combines Sorbian adaptation data with multilingual and task-oriented replay.

The complete recipe contains approximately 186.74M full tokens and 157.76M trainable tokens. The complete composition is summarized in Table 1.

Sorbian adaptation data. Our monolingual pool combines the official WMT 2022, 2025, and 2026 Sorbian resources, including Wikipedia and data released by the Sorbian Institute and Witaj Language Center (Weller-Di Marco and Fraser, 2022; Okabe et al., 2025; TUM-NLP, 2026). Because substantially more Upper Sorbian than Lower Sorbian text is available after cleaning and deduplication, we expose the Upper Sorbian pool once and the Lower Sorbian pool twice. This yields 83.17M Upper Sorbian and 27.70M Lower Sorbian tokens.

We also include 82,514 unique bilingual sentence pairs from the WMT22–26 German–Upper Sorbian, German–Lower Sorbian, and Upper–Lower Sorbian training resources. A separate directional MT replay component covers all six required translation directions. Approximately 60% of its supervised tokens are allocated to Sorbian-to-German translation, 30% to German-to-Sorbian translation, and 10% to translation between the two Sorbian languages. This allocation is a heuristic replay design rather than a separately optimized mixture. We place the largest share on Sorbian-to-German to preserve German-side generation and previously acquired multilingual capability during Sorbian-focused adaptation, retain substantial German-to-Sorbian supervision for target-language generation, and reserve the remaining portion for transfer between the two closely related Sorbian languages.

Capability replay. Because continued pre-training on a narrow Sorbian corpus may cause

¹<https://huggingface.co/worthing-zhang/sorbian-sft-v3-ck3250>

the model to over-adapt to the target language and forget capabilities acquired during multilingual pre-training, we retain English and German replay data. These replay components are intended to preserve general language modeling ability and previously learned multilingual knowledge while adapting the model to Sorbian. English replay uses FineWeb-Edu and FineWeb (Penedo et al., 2024) together with FineWiki (HuggingFaceFW, 2025). German replay uses the deu_Latn subset of FineWeb2 (Penedo et al., 2025) and German FineWiki. These components contribute 13.53M English and 9.29M German tokens.

We additionally construct Sorbian instruction-following data from 955 English instruction-response records drawn from a processed Magpie Pro pool (Xu et al., 2024). Each record is translated into both Upper and Lower Sorbian using Qwen3.7-Max via OpenRouter with temperature 0.2, JSON-formatted output, and reasoning disabled. The translation prompt requires all natural-language dialogue turns to be translated while preserving code, URLs, markup, formulas, variables, placeholders, and option labels. We reserve 50 translated records per language for validation, leaving 905 training records for each Sorbian language. These examples are used only in the mixed-CPT mixture and are repeated three times.

Mathematical replay uses the original English versions of GSM8K (Cobbe et al., 2021), level-1–4 MATH examples (Hendrycks et al., 2021), and 13 MathInstruct subsets (Yue et al., 2023).

We additionally include a lightweight task-anchor component for QA, SC, GC, and MR. These examples expose the model during CPT to the task-specific instruction and output formats that will later be used during multitask SFT: QA anchors use Sorbian question-answering formats, SC and GC anchors use controlled orthographic and grammatical corrections, and MR anchors use Sorbian problem statements with gold or deterministically verified reasoning supervision. MT is handled separately by the directional MT replay described above. The task-anchor component contributes 16.12M full sequence tokens, as shown in Table 1.

Because this stage combines ordinary next-token prediction with assistant-only supervised replay, we refer to it as mixed CPT.

2.2 Multitask SFT Data

Although CPT improves Sorbian language representations, it does not directly teach the model

Task	Example exposures	Token share
MT	180,356	47.25%
QA	78,764	3.53%
SC	58,474	9.52%
GC	78,554	12.58%
MR	37,514	27.12%
Total	433,662	100%

Table 2: Effective exposure in the final recipe multitask SFT mixture after sampling and repetition over two epochs.

to follow the heterogeneous output formats required by the WMT 2026 tasks. We therefore perform multitask supervised fine-tuning to align the adapted model with MT, QA, SC, GC, and MR objectives. Because these tasks differ substantially in data scale and output structure, we construct task-specific supervision and control their effective exposure during training.

The final SFT recipe jointly trains all five task families for two epochs. After sampling and repetition, the model receives 433,662 example exposures and approximately 15.61M supervised target tokens.

MT uses the same six-direction symmetric training pool described above. The remaining tasks use task-specific data construction designed to increase Sorbian supervision while preserving the official task format.

QA. We construct 20,000 QA examples deterministically from native Upper and Lower Sorbian Wikipedia contexts rather than translating English QA data, with 10,000 examples per language. For each language, 6,000 examples are statement-selection questions and 4,000 are Boolean-verification questions. For statement selection, the correct option is an evidence sentence from the current context, while distractors are drawn from other contexts and verified not to occur in the current context. The number of candidate options varies with difficulty from two to four. For Boolean verification, positive examples use an original evidence sentence, whereas negative examples are created by deterministically modifying a single numeric token by an offset of 1–9 and retaining only statements that do not occur in the original context. All examples preserve the corresponding Sorbian language and include the context, candidate options, answer, and supporting evidence.

SC. The final SC pool contains 59,693 accepted examples, including 31,039 synthetically corrupted

sentences and 28,654 unchanged CORRECT examples. Source sentences are drawn from Sorbian monolingual data and MT target-side text. Each corrupted example modifies exactly one token using one or two character-level operations selected from deletion, insertion, substitution, adjacent transposition, and diacritic change. The resulting pool contains 14,748 examples at character edit distance one and 16,291 at edit distance two. For unchanged examples, both the incorrect-word and corrected-word fields are set to CORRECT.

GC. Our grammar-correction data contains a balanced 40K pool, with 20K examples per language and approximately half of the examples representing already-correct sentences. Most error examples are constructed through rule-based or otherwise controlled single-token perturbations. To improve coverage of under-represented adjective-agreement errors, we replace 2,400 over-represented case-related error examples with adjective-agreement minimal pairs generated by Qwen3.7-Max, using 1,200 examples for each Sorbian language. Generation is constrained to modify exactly one whitespace-delimited token and to introduce a morphological disagreement in case, number, or gender rather than spelling or punctuation noise. Generated pairs are retained only when they satisfy the single-token constraint, have a token-core edit distance of at most three, and contain token cores of at least four characters. The LLM-generated examples therefore form part of the 40K pool rather than an additional dataset; the enhanced GC stratum used in later SFT recipes corresponds to the training-split subset of these same synthetic examples.

MR. The final MR dataset contains 18,757 unique examples: 10,363 Sorbian GSM8K examples (Cobbe et al., 2021), 3,558 Sorbian ASDiv/SVAMP examples (Miao et al., 2021; Patel et al., 2021), and 4,836 English MATH examples (Hendrycks et al., 2021).

For Sorbian examples, only the problem statement is translated. The translation is performed using gpt-5.5 with temperature 0.0, without providing answers or reasoning traces to the translation model. The prompt requires preservation of numbers, fractions, percentages, units, and symbolic expressions, followed by deterministic checks on numeric consistency, symbols, language validity, and length ratio.

GSM8K retains its original English gold rea-

soning traces, while ASDiv/SVAMP reasoning is constructed from deterministically verified equations whose results match the original answers. The MATH portion uses level-1–4 gold solutions in English. External LLMs are therefore used only for Sorbian problem translation rather than for generating mathematical reasoning supervision.

The redistributable portions of our synthetic and derived training data, together with generation prompts, filtering procedures, reconstructed processing scripts, source manifests, and license metadata, are publicly available at <https://huggingface.co/datasets/worthing-zhang/wmt26-sorbian-synthetic-data>. For examples whose upstream licenses or provenance do not permit unambiguous redistribution, we release only reproducibility metadata and processing artifacts rather than the underlying text.

2.3 Filtering and Contamination Control

After constructing the heterogeneous training mixtures, we further apply contamination filtering and quality control to reduce the risk of evaluation leakage and unintended overlap with official evaluation material. These procedures are applied throughout data construction before creating the final training views. For downstream task data, we compare source groups against available evaluation material using exact matching, normalized matching, and word 8-gram near-duplicate detection. Matched source groups are quarantined before the final production view, and no matches remain under these checks.

Held-out WMT and Sorbian validation material is also removed from the CPT monolingual pool. PluraMath (Dementieva et al., 2026), which extends PolyMath (Wang et al., 2025) to underrepresented languages and underlies the official Sorbian MR evaluation, is excluded from the complete training lineage. English mathematical sources are independently decontaminated against the corresponding English PolyMath source data.

All train/validation splits are performed at source-group level. These procedures cannot exclude semantic paraphrases or overlap with unseen hidden evaluation data; we discuss this residual risk further in the Limitations.

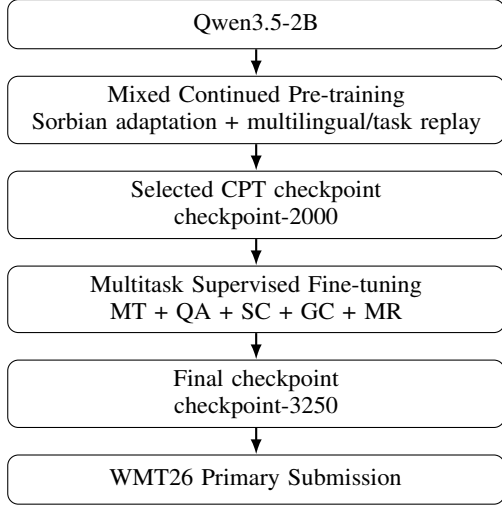


Figure 1: Overview of our two-stage adaptation pipeline. Mixed CPT strengthens Sorbian language competence while replaying multilingual and task-oriented capabilities, after which multitask SFT aligns the model with the five shared-task objectives.

3 Methodology

3.1 System Overview

Our system consists of two full-parameter adaptation stages, as illustrated in Figure 1. Mixed continued pre-training (CPT) first adapts Qwen3.5-2B to Sorbian while retaining multilingual and general task capabilities. Multitask supervised fine-tuning (SFT) then aligns the adapted model with the five shared-task objectives.

3.2 Mixed Continued Pre-training

To increase Sorbian exposure without overspecializing the model to a narrow low-resource corpus, we formulate continued pre-training as a mixed objective. Starting from Qwen3.5-2B, we jointly train on the Sorbian-focused data described in Table 1 together with multilingual and task-oriented replay.

Raw monolingual and bilingual texts are trained with the standard next-token prediction objective. Translation, instruction-following, mathematical-reasoning, and downstream-task examples instead use assistant-only supervised loss. The mixture therefore combines Sorbian language acquisition with capability replay within a single CPT stage.

We include English and German text together with directional MT supervision to reduce capability degradation during language adaptation. Lower Sorbian is also upsampled relative to its naturally available corpus size to reduce the large imbal-

ance between Upper and Lower Sorbian training resources.

The resulting CPT model is used to initialize the subsequent multitask SFT stage.

3.3 Multitask Supervised Fine-tuning

Continued pre-training improves the model’s Sorbian representations but does not directly align it with the heterogeneous input and output requirements of the five downstream tasks. We therefore perform full-parameter multitask SFT from the selected CPT checkpoint using MT, QA, SC, GC, and MR supervision jointly.

Before SFT, examples from all tasks are normalized into a common instruction–input–output schema. The instruction identifies the task, language direction when applicable, and required response format; the input contains the source sentence, context, question, or candidate options; and the output contains the canonical target response. Each record is then serialized as a single-turn conversation using the Qwen chat template, with the instruction and task input placed in the user turn and the target response placed in the assistant turn. We compute loss only on assistant tokens.

To improve instruction-following stability and permit deterministic answer extraction, the targets follow task-specific structured formats. MT targets contain only the tagged target-language translation. SC and GC targets separately identify the incorrect word and its correction. MR targets contain the verified reasoning supervision followed by a delimited final answer, while QA records preserve the candidate-answer format used for selection. The same instructions and response delimiters are retained during inference.

We train all six MT directions together with the four non-MT tasks in a single model using the final mixture reported in Table 2. Because the task datasets differ substantially in size and response length, we control effective exposure through task-specific sampling and repetition rather than sampling directly in proportion to the available raw data. Detailed construction of the task datasets is described in Section 2 and its corresponding subsections.

3.4 Inference

For the generative tasks, inference uses the same task instructions, Qwen chat serialization, and response formats as multitask SFT. This alignment avoids a train–inference prompt mismatch and al-

lows predictions to be recovered using deterministic post-processing rather than unconstrained interpretation of the complete model response.

For MT, we extract the text inside the expected target-language span and use it as the translation. For SC and GC, the incorrect-word and corrected-word fields are extracted separately and written to the corresponding submission fields. For MR, the reasoning trace is removed during post-processing and the delimited final answer is extracted for evaluation. Translation is performed independently at sentence level.

QA does not use free-form generation. We place the context, question, and candidate options in the same task template used during training, compute the conditional log-likelihood of each candidate answer, and select the highest-scoring option. The resulting task predictions are then converted to the official WMT 2026 JSONL submission fields.

4 Experimental Setup

4.1 Training Setup

Table 3 summarizes the optimization settings used for the two training stages. Both stages update all model parameters.

Setting	Mixed CPT	multitask SFT
Epochs	1	2
Max. sequence length	2,048	2,048
Learning rate	2×10^{-5}	3×10^{-6}
Warmup	3%	3%
LR schedule	cosine	cosine
Per-device batch	16	8
Gradient accumulation	2	2
Number of GPUs	8	8
Global batch	256	128

Table 3: Training configuration for mixed CPT and the final multitask SFT run.

We use ZeRO-2 (Rajbhandari et al., 2020) and FlashAttention-2 (Dao, 2024) for distributed and memory-efficient training. The final SFT run finishes after 3,388 optimization steps and takes approximately 2.75h on 8 GPUs.

The submitted checkpoint contains approximately 2.21B parameters and is evaluated in BF16 precision.

4.2 Inference Setup

For MT, SC, and GC, we disable the model’s native thinking mode and perform a single generation pass with the task-specific templates described in

Section 3.4. For MR, we enable native thinking and use temperature 1.0, top- $p = 0.95$, top- $k = 20$, and a maximum generation length of 4,096 tokens. The reasoning trace is removed before extracting the final answer. QA is evaluated through candidate conditional log-likelihood and therefore does not use generation hyperparameters.

4.3 Evaluation

We use the official WMT 2026 evaluation repository, which is based on the Language Model Evaluation Harness (Gao et al., 2023), for development evaluation and submission generation.

For MT, we report the mean chrF++ score (Popović, 2015) over the six Sorbian translation directions. BLEU is additionally reported in the official results but is not used for our internal checkpoint selection.

For QA, we use candidate-selection accuracy. For MR, we report accuracy using the official evaluation implementation. For SC and GC, our internal checkpoint-selection metric uses the corrected-form exact-match score averaged across Upper and Lower Sorbian.

For model development, we define an internal five-task selection score:

$$S_{\text{dev}} = \frac{M_{\text{MT}} + M_{\text{SC}} + M_{\text{GC}} + M_{\text{MR}} + M_{\text{QA}}}{5}, \quad (1)$$

where all task scores are represented on a 0–100 scale.

M_{MT} is the mean chrF++ score over the six MT directions. M_{SC} and M_{GC} are the mean corrected-form exact-match scores over Upper and Lower Sorbian. M_{MR} and M_{QA} are accuracy scores.

This development score is used only for model comparison and checkpoint selection. It is not identical to the official final leaderboard aggregation because the official SC and GC task scores incorporate both error detection and correction.

4.4 Checkpoint Selection

All model and checkpoint selection is performed on development data. Blind test results are not used to select the final checkpoint.

For mixed CPT, we evaluate checkpoints between steps 500 and 3000 and select checkpoint 2000 as the common initialization for subsequent SFT experiments.

For the final SFT run, we evaluate checkpoints throughout training. Checkpoint 3250 obtains the

highest value of S_{dev} among the evaluated checkpoints and is used for the primary submission.

The no-CPT ablation follows the same evaluation protocol. For that run, checkpoint 3000 obtains the highest development selection score and is used in the controlled comparison with the submitted CPT-initialized model.

5 Results and Analysis

5.1 Official Shared-task Results

Table 4 reports the official primary-submission results for the WMT 2026 Sorbian track. Detailed results by translation direction and language-specific task scores are provided in Appendix A.

As shown in Table 4, our system ranks third overall in the Sorbian track. It obtains the second-highest QA score and the third-highest score on MT, SC, GC, and MR.

Compared with the official Qwen3.5-2B baseline, our primary submission improves MT from 21.68 to 63.12 chrF++, QA from 43.39 to 57.33, SC from 6.66 to 73.13, GC from 1.34 to 62.11, and MR from 5.40 to 18.60. These correspond to absolute improvements of 41.44 points on MT, 13.94 on QA, 66.47 on SC, 60.77 on GC, and 13.20 on MR.

The leaderboard nevertheless indicates substantial remaining headroom, particularly on MT, GC, and MR. We therefore view the submission as a strong improvement over the fixed base model rather than as saturation of the Sorbian track.

5.2 Task-Mixture and Supervision Trade-offs in Multitask SFT

We examine how task allocation and supervision design affect multitask adaptation in the low-resource setting. Specifically, we compare three full-parameter SFT recipes initialized from the same selected CPT checkpoint. The comparison focuses on whether heavily increasing ordinary MT supervision improves translation, and whether a more balanced task mixture with task-aligned supervision produces a stronger overall model. Table 5 summarizes the main configuration differences and the selected checkpoint of each recipe.

As shown in Table 5, V1 provides the reference multitask mixture, whereas V2 substantially increases ordinary MT supervision to 73.94% of the supervised-token budget. Despite this increase, MT decreases from 66.01 to 65.53, an absolute reduction of 0.48 points. SC decreases from 69.55 to

64.27, and QA decreases from 69.03 to 67.47, corresponding to reductions of 5.28 and 1.56 points. GC increases from 57.47 to 59.40, but this change coincides with the introduction of refined GC supervision and therefore cannot be attributed to the increase in MT data.

The V1–V2 comparison suggests that allocating most of a fixed supervision budget to the largest task does not necessarily improve that task and may reduce the effective exposure of the remaining tasks. One possible explanation is that MT had already reached diminishing returns under the available supervision, whereas the smaller tasks were more sensitive to reductions in training exposure. Because the MR inference protocols differ, we do not use the V1 MR score or its aggregate development score to assess the effect of the SFT recipe.

V3 corresponds to the final SFT recipe used for our WMT 2026 primary submission. V3 reduces the dominance of MT and restores broader coverage of the remaining tasks. It also retains the refined GC supervision and replaces the previous MR data with the reasoning-oriented supervision described in Section 2.2. Compared with V2, V3 improves MT from 65.53 to 65.84, SC from 64.27 to 68.30, GC from 59.40 to 59.63, MR from 18.75 to 22.92, and QA from 67.47 to 70.55. These correspond to absolute gains of 0.31, 4.03, 0.23, 4.17, and 3.08 points, respectively. The overall development score consequently increases from 55.09 to 57.45, a gain of 2.36 points.

These results provide two practical lessons for low-resource multitask adaptation. Maintaining sufficient exposure for smaller tasks and improving the quality and format alignment of their supervision can be more effective than repeatedly oversampling a comparatively mature task. Since both the task mixture and the MR supervision change between V2 and V3, this comparison should be interpreted as a system-level recipe comparison rather than a controlled single-variable ablation. The complete development trajectories of the V1, V2, and V3 recipes are reported in Tables 9, 10, and 11, respectively.

5.3 Ablation Study

We conduct an ablation study to quantify the contribution of the two adaptation stages. We compare four settings: the original Qwen3.5-2B base model, mixed CPT without downstream SFT, multitask SFT applied directly to the base model, and

System	MT	QA	SC	GC	MR	Points	Rank
HeyBusan	72.75	55.92	78.03	79.52	29.00	22	1
LT3	69.57	65.05	76.11	78.32	30.40	22	1
HITSZ (ours)	63.12	57.33	73.13	62.11	18.60	16	3
HSE	61.50	46.77	63.63	50.02	5.00	9	4
Qwen3.5-2B baseline	21.68	43.39	6.66	1.34	5.40	5	5

Table 4: Official WMT 2026 Sorbian-track results for primary submissions. MT is reported in chrF++; the remaining columns use the official task scores. Bold task scores indicate the best result in each column.

Recipe	Main configuration
V1	Initial multitask mixture. It uses the initial GC and MR supervision, with non-thinking inference for MR.
V2	MT-heavy mixture, with MT accounting for 73.94% of the supervised-token budget. It uses refined GC supervision and thinking-enabled MR inference.
V3	Rebalanced mixture with MT/QA/SC/GC/MR token shares of 47.25/3.53/9.52/12.58/27.12%. It retains the refined GC supervision and replaces the previous MR data with reasoning-oriented supervision. MR uses thinking-enabled inference.

Recipe	Step	MT	SC	GC	MR	QA	S _{dev}
V1	1750	66.01	69.55	57.47	2.08 [†]	69.03	52.83 [†]
V2	3352	65.53	64.27	59.40	18.75	67.47	55.09
V3	3250	65.84	68.30	59.63	22.92	70.55	57.45

Table 5: Main configuration differences and selected development checkpoints for the three multitask SFT recipes. Percentages denote shares of supervised training tokens. [†]V1 uses non-thinking inference for MR; consequently, its MR and aggregate scores are not directly comparable with those of V2 and V3.

the complete CPT→SFT pipeline. For all comparisons, we use the same downstream evaluation protocol and select the checkpoint with the highest development selection score for each setting.

As shown in Table 6, both CPT and SFT provide substantial improvements over the original base model, but they contribute in different ways. Mixed CPT alone improves the five-task development score from 16.84 to 46.77, with particularly large gains on MT, SC, and GC. These improvements indicate that continued pre-training effectively increases Sorbian language competence and provides useful representations for language-dependent tasks.

In contrast, SFT-only achieves a score of 52.21 by directly optimizing the official downstream task formats. Compared with CPT only, SFT provides stronger improvements on MR and QA, showing that task-specific supervision is necessary for learning structured response formats and reasoning-oriented objectives.

The complete CPT→SFT pipeline obtains the best performance with a score of 57.45, outperform-

ing SFT-only by 5.24 points. The remaining gains after SFT initialization are mainly observed in MT (+6.16), SC (+14.20), and GC (+4.67), suggesting that Sorbian-oriented continued pre-training provides linguistic knowledge that cannot be fully recovered by downstream supervision alone.

The effect of CPT is task dependent. While CPT improves language-dependent tasks substantially, it does not directly solve structured reasoning evaluation: CPT-only obtains 0.00 on MR, indicating that the lightweight MR anchors used during CPT are insufficient to learn the task-specific reasoning and answer format without downstream SFT. After SFT alignment, MR performance improves to 22.92. The MR results should still be interpreted cautiously because the development set contains only 48 examples and therefore has coarse score resolution.

This ablation compares complete adaptation stages rather than individual data components. Since removing CPT changes the entire continued pre-training stage, the experiment does not isolate the contribution of individual CPT data sources. Detailed checkpoint trajectories for both CPT and no-CPT runs are provided in Appendices B and C.

5.4 Task-level Analysis

The comparisons in Tables 5 and 6 reveal two complementary sources of improvement. Mixed CPT mainly benefits tasks that depend strongly on Sorbian language knowledge: under the same multitask SFT recipe, CPT initialization improves SC from 54.10 to 68.30 (+14.20), MT from 59.68 to 65.84 (+6.16), and GC from 54.96 to 59.63 (+4.67), while producing smaller gains on QA and MR.

The SFT-stage results further show that supervision allocation affects how the adapted capability is distributed across tasks. Increasing MT to 73.94% of the supervised-token budget reduces several other task scores, whereas the rebalanced recipe achieves the strongest overall development result.

System	MT	SC	GC	MR	QA	S _{dev}
Qwen3.5-2B	22.04	8.42	0.38	4.17	49.21	16.84
CPT only	55.66	58.30	57.46	0.00	62.43	46.77
SFT only	59.68	54.10	54.96	22.92	69.38	52.21
CPT→SFT	65.84	68.30	59.63	22.92	70.55	57.45
Δ (CPT→SFT vs. Base)	+43.80	+59.88	+59.25	+18.75	+21.34	+40.61

Table 6: Ablation study of the two-stage adaptation pipeline. CPT only applies mixed continued pre-training without downstream supervised fine-tuning, SFT only directly fine-tunes the base model on downstream tasks, and CPT→SFT represents the complete system.

Taken together, these results motivate the final two-stage system. Mixed CPT supplies broad Sorbian language adaptation, while the multitask SFT recipe determines how effectively that capability is distributed across the five downstream objectives. In our setting, allocating most of the supervised-token budget to the largest task does not improve that task. Maintaining broader task coverage and using task-aligned supervision instead produces the strongest overall development result.

We treat this interpretation as specific to our training setting. The CPT comparison removes the complete adaptation stage rather than individual CPT components, and the SFT recipes are system-level comparisons rather than controlled single-variable ablations.

6 Conclusion

We present the HITSZ system for the WMT 2026 Multitask LLMs with Limited Resources Sorbian track. Our approach adapts Qwen3.5-2B through two complementary stages: mixed continued pre-training combines Sorbian adaptation data with multilingual and task-oriented replay to improve language competence, while multitask supervised fine-tuning aligns the adapted model with MT, QA, SC, GC, and MR objectives. The resulting system substantially outperforms the official Qwen3.5-2B baseline on all five tasks and ranks third overall in the Sorbian track. Our ablation studies show that downstream SFT alone cannot fully replace target-language adaptation: removing mixed CPT decreases the five-task development score by 5.24 points, mainly affecting language-dependent tasks such as SC, MT, and GC. We further find that allocating excessive supervision to the largest task does not improve translation performance and can reduce other task capabilities. These results highlight the importance of combining sufficient target-language adaptation with balanced supervision allocation when adapting compact LLMs to low-

resource multilingual multitask settings.

7 Limitations

Our experiments are primarily system-development experiments rather than exhaustive controlled studies. In particular, the V1, V2, and V3 recipes differ in more than one design choice, so their comparison supports conclusions about complete training recipes rather than the contribution of individual components. We also train one main run for each recipe and do not estimate variance across random seeds, which is especially relevant for MR because its development set contains only 48 examples.

Our development protocol also differs slightly from the official leaderboard evaluation. SC and GC checkpoint selection uses corrected-form exact match, whereas the official scores combine error detection and correction. In addition, our contamination filtering cannot completely exclude semantic overlap with unseen evaluation material, and LLM-assisted or synthetic data construction may introduce artifacts that are not separately quantified.

References

- 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*.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *International Conference on Learning Representations*.
- 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, Tsendeniya Kinfe Temesgen, Galit Bary Weisberg, and

- Alexander Fraser. 2026. [Pluramath: Extending mathematical reasoning evaluation beyond high-resource languages](#). *Preprint*, arXiv:2607.05992.
- 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. 2023. [A framework for few-shot language model evaluation](#).
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). *arXiv preprint arXiv:2103.03874*.
- HuggingFaceFW. 2025. [FineWiki](#). Hugging Face dataset.
- 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.
- Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. [A diverse corpus for evaluating and developing english math word problem solvers](#). *arXiv preprint arXiv:2106.15772*.
- 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.
- Shu Okabe, Marion Di Marco, Kathy Hämmerl, Daryna Dementieva, Lukas Edman, Marko Měškank, Anita Hendrichowa, and Alexander Fraser. 2026. [Findings of the WMT26 Shared Task: Multitask LLMs with Limited Resources](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The FineWeb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, 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](#). *arXiv 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.
- 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. [Qwen3.5-2b](#). Hugging Face model card, accessed 2026-08-18.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. [ZeRO: Memory optimizations toward training trillion parameter models](#). In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE.
- 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.
- TUM-NLP. 2026. [Multitask llms with limited resources @ wmt 2026](#). Accessed 2026-08-18.
- Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [PolyMath: Evaluating mathematical reasoning in multilingual contexts](#). *arXiv preprint arXiv:2504.18428*.
- 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](#). *arXiv preprint arXiv:2406.08464*.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen.

2023. [MAMmoTH: Building math generalist models through hybrid instruction tuning](#). *arXiv preprint arXiv:2309.05653*.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-MT technical report](#). *Preprint*, arXiv:2509.05209.

A Detailed Evaluation Results

A.1 Machine Translation by Direction

Table 7 reports the official MT results of our primary submission for all six Sorbian translation directions.

Direction	BLEU	chrF++
de→hsb	35.32	60.04
hsb→de	37.42	62.42
de→dsb	26.87	52.37
dsb→de	36.30	60.92
dsb→hsb	49.42	70.62
hsb→dsb	50.50	72.35

Table 7: Official machine translation results of our primary submission by direction.

Translation between Upper and Lower Sorbian is strongest, exceeding 70 chrF++ in both directions. German-to-Lower-Sorbian is the weakest direction at 52.37 chrF++.

A.2 Detailed QA, SC, GC, and MR Results

Table 8 reports the official Upper and Lower Sorbian results for the non-MT tasks.

Metric	HSB	DSB
QA accuracy	58.57	56.10
MR accuracy	19.20	18.00
SC detection	81.55	78.05
SC correction	66.95	65.95
GC detection	64.70	63.09
GC correction	60.80	59.85

Table 8: Detailed official results for Upper Sorbian and Lower Sorbian.

B Development Checkpoint Trajectories

B.1 SFT V1

Step	MT	SC	GC	MR	QA	S _{dev}
250	60.40	62.60	57.05	4.17	64.86	49.82
500	61.58	63.98	57.64	0.00	65.65	49.77
750	62.96	66.15	57.69	4.17	66.90	51.57
1000	65.51	69.42	57.10	6.25	68.30	53.32
1250	65.81	69.55	57.37	2.08	67.63	52.49
1500	66.09	70.10	57.04	2.08	68.20	52.70
1750	66.01	69.55	57.47	2.08	69.03	52.83
1782	66.02	69.80	57.27	0.00	67.98	52.21

Table 9: Development trajectory for the V1 SFT run. V1 MR uses non-thinking evaluation.

B.2 SFT V2

Step	MT	SC	GC	MR	QA	S _{dev}
500	60.83	60.02	55.68	14.58	64.13	51.05
1000	62.63	64.98	59.96	10.42	65.08	52.61
1500	64.42	63.13	58.87	18.75	65.46	54.12
2000	64.65	62.02	58.19	14.58	66.96	53.28
2500	65.39	63.62	57.55	16.67	67.41	54.13
3000	65.67	63.35	59.61	14.58	67.47	54.14
3352	65.53	64.27	59.40	18.75	67.47	55.09

Table 10: Development trajectory for the V2 SFT run.

B.3 SFT V3

Step	MT	SC	GC	MR	QA	S _{dev}
1500	63.02	65.70	58.42	12.50	68.10	53.55
1750	64.82	67.95	60.76	20.83	68.84	56.64
2500	65.35	68.17	60.10	16.67	69.12	55.88
3000	65.65	67.78	60.80	20.83	70.33	57.08
3250	65.84	68.30	59.63	22.92	70.55	57.45
3388	65.72	68.35	60.07	20.83	70.33	57.06

Table 11: Development trajectory for the final V3 SFT run.

C No-CPT Ablation Trajectory

Table 12 reports the complete development trajectory of the SFT-only recipe initialized directly from Qwen3.5-2B without mixed CPT. The best development score is obtained at step 3000, which is used in the controlled comparison reported in Table 6.

Step	MT	SC	GC	MR	QA	S _{dev}
250	40.47	35.43	40.84	12.50	65.74	39.00
500	45.57	43.07	46.37	14.58	66.00	43.12
750	48.69	43.25	49.04	14.58	67.78	44.67
1000	52.35	45.45	52.23	14.58	65.91	46.10
1250	53.77	48.75	51.83	16.67	69.05	48.01
1500	54.81	52.33	54.42	14.58	68.62	48.95
1750	57.33	53.12	54.78	18.75	70.60	50.92
2000	58.32	53.52	53.60	20.83	69.86	51.23
2250	59.30	52.30	53.95	20.83	69.26	51.13
2500	59.51	54.30	55.05	16.67	69.26	50.96
2750	59.60	53.12	54.42	22.92	69.54	51.92
3000	59.68	54.10	54.96	22.92	69.38	52.21
3250	59.91	53.40	54.72	22.92	69.60	52.11
3388	59.90	53.62	55.18	22.92	69.38	52.20

Table 12: Development trajectory for the SFT-only recipe initialized directly from Qwen3.5-2B without mixed CPT.