

TUMHN@WMT2026: LLMs with Limited Resources for Slavic Languages

Tsedeniya Kinfe Temesgen^{1,2} and Sinan Kurtyigit^{1,3} and Alexander Fraser^{1,2}

¹School of Computation, Information and Technology,
Technische Universität München (TUM)

²Munich Center for Machine Learning (MCML)

³Institute for Natural Language Processing, Universität Stuttgart
{tsedeniya.temesgen|sinan.kurtyigit|alexander.fraser}@tum.de

Abstract

We describe the TUMHN team’s systems for the WMT26 Shared Tasks on LLMs with Limited Resources for Slavic Languages. We participated in the Ukrainian track, covering machine translation, question answering, spell checking, grammar checking, and math reasoning tasks. We restructured our training dataset into a chat-style format and fine-tuned the Qwen3.5-2B model for the Ukrainian track in a multitask learning setting. Our model outperforms the baseline on three of the five tasks, with significant improvements on Question Answering, Spell Checking, and Grammar Checking. In Machine Translation, the model shows a slight improvement for the English–Ukrainian direction. For Spell Checking and Grammar Checking specifically, the largest gains were observed in spell or grammatical error detection as opposed to correction. In Math Reasoning, however, our model does not outperform the baseline. The fine-tuned model ¹ and training dataset ² are publicly available on Hugging Face.

1 Introduction

Large language models (LLMs) no longer just dominate NLP research; they have found their way into people’s everyday lives worldwide. Frontier models such as ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google) are widely used to assist people in both professional and private contexts, while smaller, typically open-weight models such as Llama (Meta AI), Qwen (Alibaba), and Deepseek (DeepSeek) are used for more individualized applications. Both the training and evaluation of these models are largely biased towards high-resource languages, particularly English (Xu et al., 2025). For lower-resource languages, the number

of models that support them declines quickly (Unjum et al., 2026), and even when a model supports a particular lower-resource language, research has shown that performance is better when prompted in English (Zhang et al., 2023). This leaves speakers of lower-resource languages at a disadvantage, limiting their access to this widely used technology. Closing this gap matters regardless of whether a language is high-resource, low-resource, or somewhere in between.

The WMT26 Limited Resources LLM shared task (Okabe et al., 2026) addresses this imbalance, focusing this year on three Slavic languages: Ukrainian, Upper Sorbian, and Lower Sorbian. Ukrainian, though not low-resource, sits in the mid-resource range, with roughly 40M first-language speakers, making progress here beneficial to a large population. Upper and Lower Sorbian, by contrast, are at the extreme low-resource end, both in terms of available data and speaker populations (roughly 30K and 7K speakers, respectively). The shared task evaluates LLM performance across five tasks: machine translation (MT), multiple-choice question answering (QA), spell checking (SC), grammar checking (GC), and math reasoning (MR). Submissions are restricted to Qwen3.5-2B³ (Yang et al., 2025) as the base model.

In this work, we participated in the Ukrainian track. Our approach uses parameter-efficient fine-tuning, training a single Low-Rank Adaptation (LoRA) (Hu et al., 2021) adapter on all five tasks simultaneously. We sourced training data for the five tasks from various sources, including the Ukrainian datasets provided by the shared task organizers. Ukrainian training data was not available for math reasoning, so we used English data instead, with the expectation that the model would be able to transfer the underlying task knowledge to Ukrainian. (cf. Section 2). We restructured all task data into a

¹https://huggingface.co/Tsedeniya/qwen3.5-2B-sft-wmt2026_CFT

²https://huggingface.co/datasets/Tsedeniya/Ukrainian_MultiTask_WMT26

³<https://huggingface.co/Qwen/Qwen3.5-2B>

chat-style format and fine-tuned Qwen3.5-2B for Ukrainian in a multitask learning setting, allowing a single model to handle all five tasks jointly. Our model achieved higher performance in three of five tasks. It outperforms the baseline in question answering, spell checking, and grammar checking. For the latter two, our model performs better at detecting spelling and grammatical errors than at providing corrections. In contrast, our model does not surpass the baseline in math reasoning, likely due to the difficulty of cross-lingual transfer from English-only training data. In machine translation, our model performs slightly better on English–Ukrainian translation but falls behind on Czech–Ukrainian translation.

2 Dataset and Tasks

We participated in the **Ukrainian track**, which comprises five NLP tasks that cover different language understanding and generation settings. We convert the data from each task into a unified chat-style format, allowing us to combine all tasks and train a single model jointly on all five tasks. The datasets are summarized in Table 1; this includes both the datasets provided by the shared task and those we have gathered from several external sources.

MT: This task covers two translation directions: Czech–Ukrainian (CES-UKR) and English–Ukrainian (ENG-UKR). We sampled a subset of parallel sentences for both language pairs from the Tatoeba corpus (Tiedemann, 2020).

QA: We used the data provided by the shared task organizers, which included the MMLU (Yukhymenko et al., 2026) and ZNO (Syromiatnikov et al., 2025) datasets, structured as multiple-choice question answering. We augmented this with a Ukrainian version of SQuAD (Stanford Question Answering Dataset) (Ivanyuk-Skulskiy et al., 2021), semi-automatically translated from the original English SQuAD (Rajpurkar et al., 2016), which is formulated as an extractive QA task.

GC: This task requires the model to identify grammatical errors in a given sentence and provide the corresponding correction. Each sentence contains at most one grammatical error in a single word. Some sentences may be error-free, in which case the model should leave the sentence unchanged or output *correct*. For sentences containing an error, the model is expected to identify the

erroneous word and provide its correct form. For this task, we use data from the UA-GEC dataset (Syvokon et al., 2023), which covers a range of grammatical and fluency-related errors, as well as spelling and punctuation errors.

SC: Each sentence contains at most two spelling errors within a single word. As in GC, some sentences may be error-free and should therefore be left unchanged or marked as *correct* by the model. When an error is present, the model should identify the misspelled word and provide its correct spelling. The training dataset is sourced from UA-GEC (Syvokon et al., 2023).

MR: We incorporated data from OlympiadBench (He et al., 2024), which contains mathematics, physics, and multimodal reasoning problems. From this dataset, we selected the text-only, open-ended questions from the mathematics and physics subsets. We additionally used data from MathMist (Sobhani et al., 2025), a multilingual mathematical reasoning benchmark, from which we selected the English subset, as Ukrainian was not available.

Reid and Artetxe (2023) show that parallel data facilitates cross-lingual transfer not only for translation but also for other downstream tasks. Motivated by this finding, we hypothesize that machine translation resources can similarly support the transfer of task-specific capabilities across languages. In our setting, both OlympiadBench and MathMist are in English. We therefore hypothesize that a model can acquire mathematical reasoning ability from English-language examples and transfer this ability to Ukrainian, using English–Ukrainian parallel data as an intermediate signal for cross-lingual adaptation.

In total, we collected 66.9K training samples. The datasets are not evenly balanced, with MT taking the largest proportion of the training data, followed by the SC dataset, as shown in Figure 1. This distribution largely reflects the availability of data across the different Ukrainian datasets. The datasets for each task are merged within each split and shuffled before training⁴.

3 Methods

We fine-tuned the 4-bit quantized Qwen3.5-2B model (non-thinking mode) using Low-Rank Adap-

⁴For datasets with predefined training and validation splits, we use the provided splits. When no validation split is available, we randomly divide the training data into 80% training and 20% validation sets.

Task	Source	Train	Test
MT	CES/ENG → UKR Tatoeba (Tiedemann, 2020)	32,000	317
	† MMLU (Yukhymenko et al., 2026)		
QA	† ZNO (Syromiatnikov et al., 2025)	8,596	14,793
	UA-SQuAD (Ivanyuk-Skulskiy et al., 2021)		
SC	UA-GEC (Syvokon et al., 2023)	17,016	2,000
GC	UA-GEC (Syvokon et al., 2023)	6,929	2,000
MR	MathMist (Sobhani et al., 2025)	2,417	250
	OlympiadBench (He et al., 2024)		
Total training instances		66,958	

Table 1: Datasets used for each subtask in the Ukrainian track. † indicates a dataset provided by shared task organizers.

tation (LoRA) with a rank of $r = 32$ on a single NVIDIA H100 GPU for approximately 4 hours. The model was initially trained for 3 epochs and subsequently fine-tuned for an additional 10 epochs⁵, with early stopping (patience of 3). Loss is computed only over the assistant’s response tokens. We use a zero-shot strategy with greedy decoding during both training and inference. For reproducibility, we report the hyperparameters in Appendix A, as well as the model instructions used for each task in Appendix B.

Evaluation: We report **chrF++** (Popović, 2015) as the main evaluation metric for MT and **BLEU** (Papineni et al., 2002) for comparison. For QA and MR, we report **accuracy**. Finally, for SC and GC, we report **accuracy** separately for detection and correction.

4 Results and Discussions

Table 2 shows our model’s overall performance compared with the baseline results reported by the shared-task organizers and the two other participating teams, Zolint and koshi. According to the shared-task scoring scheme, Zolint achieves the highest overall score with 19 points, followed by koshi with 12 points. TUMHN obtains 9 points, the same overall score as the baseline, placing us fourth.

Despite obtaining the same overall score, our approach and the baseline exhibit substantially different performance profiles across the individual tasks. Our approach outperforms the baseline in three of the five tasks: QA, SC, and GC. In contrast,

⁵We initially trained the model for 3 epochs and evaluated its performance on the dev set for each task. However, the results were not satisfactory; therefore, we continued training on the same dataset for an additional 10 epochs with early stopping.

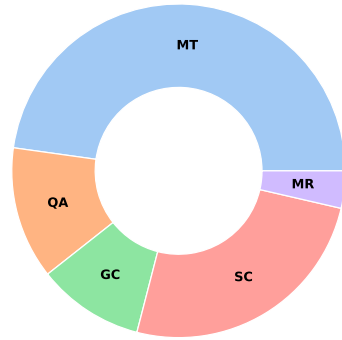


Figure 1: Distribution of training examples across tasks. All datasets are in Ukrainian except for MR, which is in English.

the baseline performs better in MT and MR (See Table 3).

MT: MT results show a mixed pattern across translation directions. On the CES-UKR subset, the baseline outperforms our model (BLEU 5.38 vs. 3.93; chrF++ 25.44 vs. 19.77). In contrast, our model performs slightly better on ENG-UKR, with BLEU 1.01 vs. 0.55 and chrF++ 12.39 vs. 11.47. Overall, our approach shows small gains on ENG-UKR but larger losses on CES-UKR, leading to a lower aggregate MT score. Moreover, the test set is constructed at the document level, whereas our model is trained at the sentence level. This discrepancy may contribute to the model’s low performance, consistent with the findings of Alabi et al. (2025), which report that models trained on sentences often struggle to generalize to longer documents. We notice incomplete translations in CES-UKR, in which the model provides short responses. This is likely due to the training data consisting of short sentences. Nevertheless, this shows an improvement over the baseline, which frequently failed to produce a meaningful translation, often refusing to translate, or generating incoherent output.

QA: Our model outperforms the baseline on QA, achieving an overall accuracy of 40.13 versus 36.33 for the baseline (+3.80 points). This improvement holds consistently across both QA domains. On UkrQA, our model reaches 35.82 compared to 30.76 for the baseline (+5.06 points), while on Ukr-MMLU it achieves 44.44 vs. 41.90 (+2.54 points).

SC: Our model shows its strongest gain over the baseline in SC, reaching an overall accuracy of 16.50 (+8.12 points). The gain comes mainly from error detection; our model scores 22.20 vs. 8.55 for the baseline, showing it is much better at recognizing misspelled word(s). However, on correction,

	MT (chrF++)	Points	QA	Points	SC	Points	GC	Points	MR	Points	Final Points	Final Rank
Zolint	49.86	4	41.15	3	66.43	4	45.18	4	21.60	4	19	1
koshi	16.80	1	46.99	4	52.50	3	30.55	3	0.40	1	12	2
baseline	18.46	3	36.33	1	8.38	1	2.18	1	20.00	3	9	4
TUMHN	16.08	1	40.13	2	16.50	2	6.80	2	4.40	2	9	4

Table 2: Ukrainian track leaderboard: best results are in **bold**.

the gap is smaller (10.80 vs. 8.20). The predominant error pattern observed was that the model flags a word as erroneous but predicts the same word as the corrected output, leaving the erroneous word unchanged.

GC: We achieved an accuracy score of 6.80 compared to 2.18 for the baseline (+4.62 points). Looking more closely, this advantage comes mainly from detecting grammatically incorrect instances, where our model reaches 10.35 vs. 2.70 for the baseline, whereas the gap on corrected instances is narrower (3.25 vs. 1.65). Similar error patterns are observed here as in SC, suggesting that the model is more effective at flagging errors than at generating accurate corrections.

MR: MR shows the largest performance gap between our approach and the baseline, with an accuracy score of 4.40 versus 20.00 for the baseline. This result shows whether parallel data can support cross-lingual transfer of task capabilities. Our training data combines English math data with English-Ukrainian parallel data and the aforementioned Ukrainian data; therefore, we expected this combination to help transfer mathematical reasoning ability to Ukrainian. While parallel data has been shown to support cross-lingual transfer under certain training setups (Reid and Artetxe, 2023), another study shows that MT training does not necessarily improve cross-lingual representations for downstream tasks (Ji et al., 2024). Our results align with this latter finding: despite joint training on English MR data and parallel data, our approach substantially underperforms the baseline on Ukrainian MR. Furthermore, our MT dataset is derived from Tatoeba, which contains simple sentences. This could explain the lower performance on MR, which is a more complex task that requires a higher level of linguistic proficiency.

5 Conclusion

We presented our submission to the Ukrainian track of the WMT2026 Shared Task on LLMs with Limited Resources for Slavic Languages. Our approach

Task	Subset	Metric	TUMHN	Baseline
MT	ces-ukr	BLEU	3.93	5.38
	ces-ukr	chrF++	19.77	25.44
	eng-ukr	BLEU	1.01	0.55
	eng-ukr	chrF++	12.39	11.47
	all	chrF++	16.08	18.46
QA	ukrqa	ACC	35.82	30.76
	ukrmmlu	ACC	44.44	41.90
	all	ACC	40.13	36.33
SC	ukrsc	ACC-D	22.20	8.55
		ACC-C	10.80	8.20
	all	ACC	16.50	8.38
GC	ukrgc	ACC-D	10.35	2.70
		ACC-C	3.25	1.65
	all	ACC	6.80	2.18
MR	ukrmr	ACC	4.40	20.00

Table 3: Results for the Ukrainian track: We compare TUMHN (our approach) vs. the baseline results reported by the organizers. BLEU, chrF++, and exact-match accuracy are reported. Best scores are presented in **bold**. $ACC - D$ and $ACC - C$ represent accuracy scores for detection and correction, respectively.

consisted of fine-tuning Qwen3.5 2B model jointly on five tasks: machine translation, math reasoning, spell checking, grammar checking, and question answering.

Our system ranked fourth among participants, tied with the baseline. Looking at the individual tasks, we outperformed the baseline on QA, SC, and GC, with the largest gains on SC, while falling short on MT and MR. We hypothesized that task knowledge learned from English MR data could transfer to Ukrainian. However, our results do not support this. Performance on MR remains below the baseline, suggesting that cross-lingual transfer alone was insufficient to compensate for the lack of in-language training data.

Overall, we view these results as encouraging. Despite using a modest amount of data and a comparatively short training time, we improved over the baseline on three of the five tasks, demonstrating that this joint fine-tuning approach is a viable strategy for mid-resource languages like Ukrainian under limited compute and data budgets.

6 Limitations

Our submission has several limitations that we leave for future work:

- **Single training strategy.** We have only explored joint fine-tuning on all five tasks simultaneously, without comparing against alternative strategies such as sequential or task-specific fine-tuning. Additionally, we only evaluated our model on a single random seed.
- **No task interaction analysis.** We did not analyze how the tasks influence one another during joint training. Some tasks may contribute more positively to overall model capability than others, or even have a negative effect; identifying and prioritizing such beneficial tasks, e.g. through additional data for those specifically, is a promising direction for future work.
- **No translation for MR.** Due to the lack of Ukrainian data for MR, we relied on English data and cross-lingual transfer. We did not explore machine-translating the English MR data into Ukrainian before training, which might have been a more effective approach.

7 Acknowledgement

This work was co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) 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.

References

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo

Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. *AFRIDOC-MT: Document-level MT corpus for African languages*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27770–27806, Suzhou, China. Association for Computational Linguistics.

Alibaba. Qwen. <https://qwenlm.github.io>. Accessed: 2026-08-17.

Anthropic. Claude. <https://www.anthropic.com/claude>. Accessed: 2026-08-17.

DeepSeek. DeepSeek. <https://www.deepseek.com>. Accessed: 2026-08-17.

Google. Gemini. <https://gemini.google.com>. Accessed: 2026-08-17.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. Preprint, arXiv:2106.09685.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. *ua₄datasets : a collection of ukrainian language datasets*.

Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024. *Can machine translation bridge multilingual pretraining and cross-lingual transfer learning?* In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2809–2818, Torino, Italia. ELRA and ICCL.

Meta AI. Llama. <https://www.llama.com>. Accessed: 2026-08-17.

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.

OpenAI. ChatGPT. <https://openai.com/chatgpt>. Accessed: 2026-08-17.

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.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Machel Reid and Mikel Artetxe. 2023. [On the role of parallel data in cross-lingual transfer learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5999–6006, Toronto, Canada. Association for Computational Linguistics.
- Mahbub E Sobhani, Md. Faiyaz Abdullah Sayeedi, Tasnim Mohiuddin, Md Mofijul Islam, and Swakkhar Shatabda. 2025. [Mathmist: A parallel multilingual benchmark dataset for mathematical problem solving and reasoning](#). *Preprint*, arXiv:2510.14305.
- Mykyta Syromiatnikov, Victoria Ruvinskaya, and Anastasiya Troynina. 2025. Zno-eval: Benchmarking reasoning capabilities of large language models in ukrainian. *arXiv preprint arXiv:2501.06715*.
- 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.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Naveed Unjum, Stephanie Evert, Kengatharaiyer Sarveswaran, Ruvan Weerasinghe, and Nevidu Jayatileke. 2026. [Llms for low-resource languages: A survey](#).
- Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. [A survey on multilingual large language models: corpora, alignment, and bias](#). *Frontiers of Computer Science*, 19(11).
- 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.
- Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2026. [Recovered in translation: Efficient pipeline for automated translation of benchmarks and datasets](#). *Preprint*, arXiv:2602.22207.
- Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. [Don’t trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

A Hyperparameter

Table 4 below summarizes the hyperparameters that we used for fine-tuning; We used greedy decoding in non-thinking mode, with all other decoding parameters kept at their default values in both training and inference.

Hyperparameter	Value
Learning rate	2e-5
Batch size	16
Warmup steps	20
Epochs	[†] 13
Optimizer	AdamW 8Bit
Seed	3407
Temp.	1.0

Table 4: Hyperparameters used for fine-tuning. [†] The model was initially fine-tuned for 3 epochs and continued fine-tuning for 10 more epochs with early stopping; therefore, in total it was fine-tuned for 13 epochs.

B Instructions

Table 5 lists the exact instruction used for each of the five tasks, including the expected output format.

Task	Instructions
MT	Translate the following English/Czech text to Ukrainian. Put your output in this format <ukr>Ukrainian translation </ukr>.
QA	Given a multiple-choice question, select the correct option from the listed alternatives (0–3). Respond ONLY with the number corresponding to the correct answer enclosed in <answer> and </answer> tags.
SC	The following sentence contains at most one misspelled word. Identify the misspelled word, if it exists, and place it between <wrong>...</wrong> tags, with the correct spelling in <corrected>...</corrected> tags. If there is no error, use the literal string CORRECT in both tags.
GC	The following sentence contains at most one grammatical error. If an error exists, place the incorrect word inside <wrong>...</wrong> and the corrected word inside <corrected>...</corrected>. If there is no error, output CORRECT in both tags.
MR	Answer this mathematical reasoning question. Provide your final answer using appropriate LaTeX formatting when needed, following the template <answer>...</answer>.

Table 5: Instructions used to prompt the model for each of the five tasks during joint training. Each instruction specifies both the task and the expected output format, including the tags used to delimit the model’s answer.