JGU Mainz's Submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA

Hossain Shaikh Saadi,[1] **Minh Duc Bui,**[1]
Mario Sanz-Guerrero,[1] and **Katharina von der Wense**[1,2]
[1]Johannes Gutenberg University Mainz, Germany
[2]University of Colorado Boulder, USA

Abstract

This paper presents the JGU Mainz submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: Machine Translation and Question Answering, focusing on Ukrainian, Upper Sorbian, and Lower Sorbian. For each language, we jointly finetune a Qwen2.5-3B-Instruct model for both tasks with parameter-efficient finetuning. Our pipeline integrates additional translation and multiple-choice question answering (QA) data. For Ukrainian QA, we further use retrieval-augmented generation. We also apply ensembling for QA in Upper and Lower Sorbian. Experiments show that our models outperform the baseline on both tasks.

1 Introduction

While large language models (LLMs) are strong multitask learners for high-resource languages such as English, this is not the case for smaller LLMs and languages with limited data. In this setting, a trade-off presents between the performance on different tasks. The *WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA*[1] focuses on the development of relatively small LLMs ($\leq$3B parameters) that are capable of performing both machine translation (MT) and multiple-choice question answering (QA) in Slavic languages with limited amounts of data. Three languages – a mid-resource language, Ukrainian (UK), and two severely low-resource languages, Upper Sorbian (HSB) and Lower Sorbian (DSB) – are targeted, and only Qwen2.5 models with 0.5B, 1.5B, or 3B parameters are permitted. The MT source languages are Czech (CS) and English (EN) for Ukrainian as well as German (DE) for DSB and HSB.

Our proposed approach consists of a Qwen2.5-3B-Instruct model (Qwen et al., 2025), which is inherently multilingual and which we jointly finetune on both MT and QA data, combining the provided resources with additional datasets we curate ourselves. For DSB and HSB MT, we enhance the training data using synthetic data generated through back-translation. We further add an additional parallel dataset for DSB. For QA, we add a total of 16 high-quality English MCQ datasets. All QA datasets are enhanced via automatic translation such that they are bilingual (English and the target language). For Ukrainian QA, we incorporate retrieval-augmented generation (RAG) using domain-relevant Wikipedia pages and 10 books related to the subjects of the provided Ukrainian MCQ dataset.

At inference time, we use similarity-based few-shot in-context learning (ICL) for MT. For QA, we permute the order of the answer options, and average the probabilities for all options, to increase robustness against answer ordering biases (Pezeshkpour and Hruschka, 2024).

While not winning for any task–language combination, our primary submission consistently outperforms the baseline on all tasks, demonstrating the effectiveness of our approach. For DE–DSB translation, ChrF++ improves by over 55 points, while DE–HSB translation sees gains of over 65 points, reflecting substantial quality improvements. On DSB QA, accuracy increases by up to 12.34 percentage points, and, on HSB QA, accuracy increases by 10.27 points. For CS–UK and EN–UK MT, ChrF++ improves by 4.61 and 2.7, respectively, while, on Ukrainian QA, our submission outperforms the baseline by 4.66 accuracy points.

2 Data

2.1 Provided Data

Ukrainian For EN–UK and CS–UK MT, no training data are provided. Only development sets are given, containing 6,263 CS–UK and 5,108 EN–

[1] https://www2.statmt.org/wmt25/limited-resources-slavic-llm.html

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 909–915
November 8–9, 2025. ©2025 Association for Computational Linguistics

UK parallel sentences.

The UK QA data are curated from the External Independent Evaluation (3HO/ZNO), an exam for admission to Ukrainian universities. The dataset comprises a training set of 2,450 questions, a development set of 613 questions, and a test set of 751 questions. These questions cover three topics – Ukrainian History, Ukrainian Language, and Ukrainian Literature – and test both domain knowledge as well as reading comprehension.

Upper and Lower Sorbian MT For DE-to-DSB MT, 171k translation pairs are provided as training data. In addition, a 4,000-pair development set is also provided for system validation and evaluation. Some monolingual sentences, approximately 10k, are also provided along with the translation data.

For DE-to-HSB MT, 187k training translation pairs and a 4,000-pair development set are provided. 300k monolingual sentences are further available for model enhancement.

For both DSB and HSB, MCQ datasets are curated by the Witaj-Sprachzentrum. The questions are similar to the CEFR framework (A1 to C1), which follows the language certificate examinations. The difficulty of the questions ranges from simple true/false to complex multiple-choice formats with two to sixteen answer options. The development set contains 158 A1 to B2 level questions, while the test set for both languages has 205 questions for A1 to C1. An overview of the provided datasets is shown in Table 1.

2.2 Additional Data

Upper and Lower Sorbian MT In order to enhance our DE–DSB/HSB translation systems, we incorporate additional parallel data. For DSB, we translate the provided 10k monolingual examples into German. Then we create 10k additional translation pairs using the translated German sentences and the monolingual DSB sentences. Since 10k is a small amount, we additionally use 24k DE–DSB sentences from the Tatoeba bilingual dataset.[2] In turn, for HSB, we translate the first 100k monolingual sentences from the provided 300k sentences into German and create 100k additional translation pairs. Due to the substantial time required for translation, we translate only 100k sentences. To translate the DSB and HSB sentences into German, we first finetune two separate Qwen2.5-3B-Instruct models, one for HSB–DE and one for DSB–DE, using the respective provided parallel translation datasets. These models are finetuned by applying LoRA on all projection layers of the model.

Upper Sorbian, Lower Sorbian, and Ukrainian QA For QA in DSB, HSB, and Ukrainian, we select 16 English MCQA datasets, namely: (English) Global MMLU (Singh et al., 2025), CommonsenseQA (Talmor et al., 2018), ARC (Clark et al., 2018), Race (Lai et al., 2017), Dream (Sun et al., 2019), PIQA (Bisk et al., 2019), HellaSwag (Zellers et al., 2019), SCIQ (Johannes Welbl, 2017), MedMCQA (Pal et al., 2022), LogicQA (Liu et al., 2020), Quail (Rogers et al., 2020), SocialIQA (Sap et al., 2019), CosmosQA (Huang et al., 2019), OpenbookQA (Mihaylov et al., 2018), QASC (Khot et al., 2020), BoolQ (Clark et al., 2019). From these datasets, we sample up to 10k questions from each available split (training, development, and test), resulting in approximately 200k English MCQs. In order to translate the English MCQs into DSB and HSB, we first use `googletrans`[3] to translate the German sentences of the provided DE–DSB and DE–HSB translation examples into English, in order to create English–DSB/HSB translation pairs. Second, using this data, we finetune two separate Qwen2.5-3B-Instruct models on EN–DSB and EN–HSB MT. We use these two models to translate the English MCQs into DSB and HSB. For Ukrainian, we also translate the 200k English MCQs using `googletrans` directly, since Ukrainian is supported by Google Translate.

Ukrainian MT For CS–UK MT, we collect training data from OpenSubtitles (OPUS, 2003; Lison and Tiedemann, 2016), NeuTED (Qi et al., 2018), KDE4 (OPUS, 2003), and ELRC UKR Acts (ELRC, 2022). For EN–UK, we use OpenSubtitles (OPUS, 2003; Lison and Tiedemann, 2016), NeuTED (Qi et al., 2018), ELRC UKR Acts (Qi et al., 2018), and Multi30k (Saichyshyna et al., 2023). Across these sources, there are nearly 7 million sentence pairs per direction.

To reduce the number of sentence pairs, we apply a similarity-based retrieval method, using the provided development sets for CS–UK and EN–UK as the reference datasets. We embed each Ukrainian sentence from this dataset by performing mean-pooling over the last hidden states of Qwen2.5-3B-Instruct token outputs. For each sentence of

[2] https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/README-v2023-09-26.md

[3] https://github.com/ssut/py-googletrans

Task	Training Data	Dev Set	Test Set	Notes
EN→UK Translation	None	5,108 pairs	—	Only dev set available
CS→UK Translation	None	6,263 pairs	—	Only dev set available
Ukrainian QA (MCQs)	2,450 questions	613 questions	751 questions	From ZNO exam; covers History, Language & Literature; tests knowledge and comprehension
DE→DSB Translation	171k pairs	4,000 pairs	—	∼10k monolingual sentences provided
DE→HSB Translation	187k pairs	4,000 pairs	—	∼300k monolingual sentences provided
DSB QA (MCQs)	—	158 A1–B2 questions	205 A1–C1 questions	From Witaj-Sprachzentrum; CEFR-style; difficulty from true/false to multiple-choice (2–16 options)
HSB QA (MCQs)	—	158 A1–B2 questions	205 A1–C1 questions	Same as DSB QA

Table 1: Summary of the provided datasets for MT and QA.

the reference dataset we retrieve the 75 most similar Ukrainian sentences from the collected pool of translation data along with the associated sentence in CS or EN and aggregate them. We then deduplicate the aggregated set to retain only unique translation pairs. Overall, we get 321k and 251k CS–UK and, respectively, EN–UK translation pairs for training.

2.3 Data for Retrieval-augmented Generation

For the Ukrainian QA task, we employ retrieval-augmented generation (RAG) using pages from Wikipedia and 10 books on Ukrainian history, language, and literature (the same sources used by the winning team of the UNLP 2024 Shared Task (Boros et al., 2024)). We extract about 30k pages using the `wikipediaapi`[4] library, setting `max_depth = 2` to include relevant subcategories from the Ukrainian history, language, and literature categories.

3 Models and Algorithms

3.1 Qwen: The Underlying LLM

All our models are LoRA (Hu et al., 2021)-finetuned Qwen2.5-3B-Instruct models, and, thus, satisfy the shared task's parameter constraint. We use one model per language for all tasks.

3.2 Finetuning

Finetuning on MT and General QA Data For DSB and HSB, we combine the provided MT data, additional translations created by us, and translated MCQs to finetune Qwen2.5-3B-Instruct with LoRA (Hu et al., 2021) applied to all projection layers. For DSB/HSB, we use both the English and the translated version of each MCQ, in this format:

MCQ Prompt for DSB/HSB during training

```
en_context (if any)
dsb/hsb_context (if any)

Question: {en_question}
Question: {dsb/hsb_question}

Possible Answers: {en_possible_answers}
Possible Answers: {dsb/hsb_possible_ans}

Answer: {answer}
```

To extract the model's predicted answer for QA, we end the prompt with "`Answer:`" and compute the next-token probabilities for each option label. The answer is then taken as the label with the highest probability. Following Sanz-Guerrero et al. (2025), we evaluate the probabilities of the "␣X" tokens[5] (i.e., tokens formed by the preceding space *together* with the option label), as this approach has been shown to yield better performance and calibration.

In order to use this prompt at inference time, we need to translate the provided DSB/HSB questions into English. For, first we finetuned two separate translation models, one for DSB–EN and another for HSB–EN, using the same dataset we use for EN–HSB and EN–HSB model finetuning, but this time in the opposite direction. These two translation models are also based on Qwen2.5-3B-Instruct and trained with LoRA. During the development

[4] https://github.com/martin-majlis/Wikipedia-API

[5] Where "X" denotes one of the option labels.

phase, we observe that for both DSB and HSB QA, alphabetic option labels lead to better results than numeric labels due to label bias (Zheng et al., 2024). So, for training and evaluation, we use alphabetic option labels.

The prompt for MT is of the following format:

MT Prompt during Training

```
Translate this German sentence into Upper
Sorbian. Put it in this format: <hsb> {Upper
Sorbian translation} </hsb>
<de> {German Sentence} </de>
```

For Lower Sorbian, "Upper" is replaced by "Lower" and <hsb> and </hsb> are replaced by <dsb> and </dsb>.

For Ukrainian (UK), we train the model on MT and QA data described in Section 2.2.

We apply the default chat template for Qwen2.5 and train on complete instructions (system + user + assistant), as described in Shi et al. (2024). We use LoRA (Hu et al., 2021) for 10 epochs with an initial learning rate of $1e-4$ and a linear learning rate scheduler. We save the model checkpoint after every epoch. At the end of training, we select the model with the highest BLEU (Post, 2018; Papineni et al., 2002) score, i.e., we select the model using the MT development set.

3.3 Final Finetuning on MT and In-domain QA Data for DSB and HSB

After initial finetuning of the model, we conduct a second round of finetuning. This final model finetuning is also performed by applying LoRA to all projection layers. We follow the same procedure for DSB and HSB. The provided QA development sets for DSB and HSB contain a total of 158 questions each, from language difficulty levels A1-B2. In this second stage of finetuning, we use these 158 in-domain QA examples and the first 3k translation pairs used in the initial finetuning process. To mitigate data scarcity, we apply oversampling: each QA item is repeated five times. Then, we shuffle the MT and the oversampled QA set and finetune the first finetuned model again to improve domain alignment for QA. We add the 3k translation pairs to avoid catastrophic forgetting of the MT capability of the models. The final models are used for both MT and QA during evaluation. As our dataset for the second round of finetuning is small (approximately 3.75k MT and QA examples), we tune the learning rate, searching over $1e-4, 1e-5, 1e-6$, and $1e-7$. Four separate models are trained with these learning rates exactly for one epoch. In this learning rate searching process, we exclude the questions from B2 during finetuning. We select the best learning rate based on performance for both QA (56 questions of B2 level) and MT (the first 400 samples of the 4k dev set). We follow this approach for both languages. After this experiment, we chose $1e-4$ for DSB and $1e-6$ for HSB. Then, we finetune the initial models with these learning rates for two epochs on approximately 3.75k instructions, including the questions from B2, performing no validation.

3.4 Averaging Probabilities

During QA evaluation for DSB and HSB, we generate multiple responses for each question by shuffling the order of answer options. We perform this step to mitigate positional bias (Pezeshkpour and Hruschka, 2024) as much as possible. For questions with 2–3 options, we use all permutations, which are 2 and 6, respectively. For those with more than three options, we randomly sample 20 unique answer option orders. We compute the probability distribution over the answer options under the model for each order and average them; we select the option with the highest average likelihood as the final answer.

3.5 Retrieval-augmented Generation for Ukrainian QA

We segment the retrieved pages (see Section 2.3) into chunks of 512 characters with an overlap of 64 characters and embed them using Qwen2.5-3B-Instruct. For each chunk, mean pooling is applied over the token representations obtained from the last hidden states. Embeddings are stored in two separate `ChromaDB` indexes: one for history and another for language and literature. We make embedding of each question following the same way we apply for page chunks, mean pooling over all token representations. Using the subject indicated for each question, we conduct a search in the corresponding subject-specific index. At inference time, we retrieve the 5 most relevant chunks and use them as context alongside the question.

3.6 Few-shot In-context Learning for MT

For MT, we employ few-shot in-context learning using similarity-based retrieval, following Zebaze et al. (2025). For DSB/HSB, we embed each test-set German source sentence and retrieve the 5 most

Model	ChrF++	
	DSB	HSB
Qwen2.5-3B-Instruct + LoRA(S2)(P)	66.6	77.6
Qwen2.5-3B-Instruct + LoRA(S1)	67.5	77.5
Baseline	12.21	11.87

Table 2: ChrF++ scores for DSB and HSB. *LoRA(S1)* = one round of LoRA finetuning; *LoRA(S2)* = two rounds of LoRA finetuning. *P* indicates our primary submission.

similar sentences from the development set, along with their translations. For Ukrainian, we use Ukrainian sentences for embedding generation and retrieval.

4 Results and Discussion

MT for Lower and Upper Sorbian Table 2 shows that our proposed approach yields a significant improvement over the baseline. The baseline system achieves only 12.21 ChrF++ for DSB and 11.87 ChrF++ for HSB. The first round of LoRA finetuning, indicated by *S1*, already increases ChrF++ to 67.5 and 77.5 for DE–DSB and DE–HSB, respectively. The goal for our second round of finetuning (*S2*), is to adapt the model with in-domain QA data, while retaining the model's MT capability as much as possible. For HSB, *S2* slightly improves over S1 (77.6 vs. 77.5), but, for DSB, performance drops slightly, from 67.5 to 66.6.

Model	Accuracy (%)	
	DSB	HSB
Qwen2.5-3B-Instruct + NO-FT	54.3	57.1
Qwen2.5-3B-Instruct + LoRA(S1)	48.3	50.0
Qwen2.5 + LoRA(S1) Avg.	50.7	54.3
Qwen2.5-3B-Instruct + LoRA(S2)	48.3	50.5
Qwen2.5 + LoRA(S2) Avg. (P)	51.7	55.2
Baseline	45.85	42.86

Table 3: QA accuracy scores (A1–C1) for DSB and HSB. *LoRA(S1)* = one round of LoRA finetuning; *LoRA(S2)* = two rounds of LoRA finetuning; *Avg.* = average over multiple option orders; *NO-FT* = no finetuning, i.e., direct use of Qwen2.5-3B-Instruct. *P* indicates our primary submission.

QA for Lower and Upper Sorbian For DSB and HSB QA (Table 3), the baseline accuracies of 45.85% (DSB) and 42.86% (HSB) are surpassed by all finetuned variants. Interestingly, the non-finetuned Qwen2.5-3B-Instruct model outperforms the baseline substantially, particularly for HSB (+14.24 accuracy). However, LoRA finetuning (S1 and S2) slightly reduces overall accuracy compared to the non-finetuned model, likely due to the trade-off introduced by joint MT and QA finetuning. Averaging over the results for different answer option orders improves accuracy after both rounds of finetuning (S1 and S2). It helps more after the second round of finetuning, reaching 51.7% for DSB and 55.2% for HSB. The improvements over non-averaged results demonstrate that this straightforward method is effective for low-resource QA.

Model	ChrF++	
	CS–UK	EN–UK
Qwen2.5-3B-Instruct + LoRA	8.09	3.10
Baseline	3.48	0.40

Table 4: ChrF++ scores for CS–UK and EN–UK MT.

Model	Accuracy (%)
Qwen2.5-3B-Instruct + LoRA + RAG	35.82
Baseline	31.16

Table 5: Accuracy scores for Ukrainian QA. The shown model is our primary submission.

MT for Ukrainian The Ukrainian MT tasks (Table 4) are challenging due to the lack of training data provided by the shared task. The baseline ChrF++ scores of 3.48 (CS–UK) and 0.40 (EN–UK) reflect the difficulty level. By retrieving and curating translation data via similarity search and then finetuning a Qwen2.5-3B-Instruct model, our system achieves slight improvements over the baseline for both CS–UK (8.09) and EN–UK (3.10). Though performance increases, the improvement is not as big as for the DE–DSB/HSB MT tasks. A possible reason for this is a mismatch between the training, development, and test sets: we train our models on sentences, but the test set consists of large documents and lengthy conversations.

QA for Ukrainian For Ukrainian QA (Table 5), our proposed model, based on finetuning jointly on MT and QA in combination with RAG, improves over the baseline: 31.16% vs. 35.82%. This gain is smaller than for DSB and HSB QA, which we attribute to two factors: First, the QA dataset for Ukrainian requires some factual knowledge regard-

ing the Ukrainian language, history, and literature, which makes the task harder. Second, most LLMs are underexposed to the Cyrillic script, resulting in weaker tokenization, over-splitting of words, and a decreased quality of token representations (Boros et al., 2024). This results in poor-quality embeddings for Ukrainian sentences. As a result, retrieval quality degrades: semantically close passages are missed or under-ranked, and the injected context is less helpful.

5 Conclusion

In this paper, we present JGU Mainz's submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages, addressing MT and QA in Ukrainian, Upper Sorbian, and Lower Sorbian. Our approach combines parameter-efficient finetuning of Qwen2.5-3B-Instruct with training data augmentation and RAG for Ukrainian QA. Our primary submissions outperform the provided baselines for all languages and tasks, achieving substantial ChrF++ gains for DE—DSB and DE-–HSB MT, as well as slight improvements for CS—UK and EN-–UK MT. For QA, averaging over order options increases accuracy for both DSB and HSB, while, for Ukrainian, we achieve moderate gains through RAG.

Acknowledgments

This work was supported by the Carl Zeiss Foundation through the TOPML and MAINCE projects (grant numbers P2021-02-014 and P2022-08-009I).

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641.

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 75–82, Torino, Italia. ELRA and ICCL.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

ELRC. 2022. ELRC Ukrainian Acts. https://elrc-share.eu/repository/browse/eu-acts-in-ukrainian/71205868ae7011ec9c1a00155d026706d86232eb1bba43b691bdb6e8a8ec3ccf/. [Online; accessed 05-August-2025].

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

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

OPUS. 2003. OPS Website. https://opus.nlpl.eu/. [Online; accessed 05-August-2025].

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

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.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. 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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. 2025. Mind the gap: A closer look at tokenization for multiple-choice question answering with llms. *Preprint*, arXiv:2509.15020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. *Preprint*, arXiv:2405.14394.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge dataset and models for dialogue-based reading comprehension. *Preprint*, arXiv:1902.00164.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.