PSK at WMT 2026 MIST: Task-Specialized QLoRA Adapters for Multilingual Summarization and Question Answering

Srikar Kashyap Pulipaka
Independent Researcher
srikar.kashyap@gmail.com

Abstract

We describe the PSK submission to the WMT 2026 Multilingual Instruction Shared Task. Our system uses the 3.35B-parameter Tiny Aya Global model with three QLoRA adapters, one for each task. The adapters are trained on multilingual document–summary pairs, passage-based question answering, and filtered standalone question answering. The summarization data also includes scientific papers with their author-written abstracts. On our held-out split, the context and summarization adapters perform better than our multitask adapter, which was trained only on data supplied by the organizers. Results for open QA are mixed and vary with answer length and evaluation method. We therefore submit three systems with the same context and summarization adapters but different open-QA adapters.

1 Introduction

The WMT 2026 Multilingual Instruction Shared Task (MIST) evaluates models under a 10B-parameter limit on three tasks: context-based question answering, summarization from a document in language X to language Y, and open-ended generation (WMT 2026 MIST Organizers, 2026). The test set covers 24 languages and contains both same-language and cross-lingual generation. MIST follows the broader multilingual evaluation effort introduced in WMT 2025, which found substantial variation across tasks and languages and also showed that automatic metrics are imperfect for open generation (Kocmi et al., 2025).

We use one multilingual backbone with three task-specific adapters. Tiny Aya Global (Salamanca et al., 2026) is first adapted to the provided data and then continued separately for each task. This design preserves a single 3.35B backbone while allowing the supervision and decoding policy to follow the output structure of each task.

2 Task and System Overview

MIST provides a task label at inference time. We use it to select one of three LoRA adapters and the corresponding decoding configuration (Figure 1); there is no language-specific routing.

2.1 Base Model and Initial Adaptation

We use `CohereLabs/tiny-aya-global`, a 3.35B-parameter model covering 70 languages (Salamanca et al., 2026). Its 8K context window supports the longer task inputs. An initial multitask adapter is trained on 20,293 provided examples using a deterministic 85/15 split; the resulting 3,626 held-out examples are reused for model selection. The task-specific adapters are initialized from this model.

3 Training Data

Table 1 summarizes the task-specific training data. Validation examples and exact test prompts are excluded from training.

3.1 Summarization

The 12,000-example mixture is divided evenly between general and scientific text. General examples combine provided data with CrossSum (Bhattacharjee et al., 2023), WikiLingua (Ladhak et al., 2020), and UPDESH (Chitale et al., 2026). The scientific portion uses full ACL Anthology papers (Bird et al., 2008) with author-written abstracts. No Language Left Behind (NLLB) (NLLB Team, 2024) supplies non-English targets, which are filtered for language, length, repetition, and semantic consistency. We remove test overlaps and placeholder-heavy papers. We also screen the provided data for likely misalignment by comparing each summary with chunks of its source document and discarding a pair only when both its best LaBSE similarity and best chrF score fall below fixed thresholds.

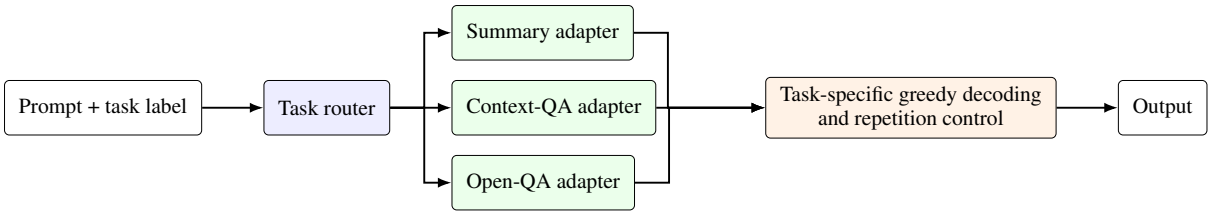

Figure 1: The three routes share one frozen Tiny Aya Global backbone. The known task label selects an adapter and its decoding policy; language does not change the route.

Adapter	Rows	Context	Main sources and construction
Multitask adapter	20,293	8,192	Provided MIST sample training split across all three tasks
Summarization	12,000	8,192	6,000 general examples from the provided data, CrossSum, WikiLingua, UPDESH, and targeted language pools; 6,000 ACL papers paired with author abstracts
Context QA	8,512	4,096	Belebele, answerable TyDi QA, MLQA, MCIF, UPDESH cultural multihop data, and small Czech/Yoruba Aya pools
Open QA	6,400	2,048	Provided open-QA examples plus filtered Aya and WMT25 MIST examples; includes Hindi questions and answers translated into Bhojpuri

Table 1: Supervised data used by the selected task adapters. Each is initialized from the multitask adapter.

3.2 Context Question Answering

The context-QA training set combines Belebele (Bandarkar et al., 2024), answerable TyDi QA (Clark et al., 2020), MLQA (Lewis et al., 2020), MCIF (Papi et al., 2026), UPDESH, and small Czech and Yoruba Aya subsets. MLQA supplies parallel and cross-lingual examples.

3.3 Open-Ended Question Answering

Open QA combines provided examples, WMT25 MIST questions, and the Aya Dataset and Collection (Singh et al., 2024). We retain short factual and explanatory QA while excluding passage-dependent, multiple-choice, creative-writing, code, translation, and continuation tasks. We also translate a small set of Hindi questions and answers into Bhojpuri with NLLB.

4 Training and Inference

We use QLoRA (Dettmers et al., 2023) with 4-bit NF4 quantization and BF16 computation. LoRA (Hu et al., 2022) is applied to the attention and feed-forward projections with rank 16, alpha 32, and dropout 0.05. Each adapter is trained for one epoch with effective batch size 16; learning rates are 1×10^{-4} for summarization, 8×10^{-5} for open QA, and 2×10^{-4} for context QA.

Inference uses greedy decoding without sampling. After observing repetitive outputs on the development set, we add repetition penalties of 1.05 for open QA and 1.03 for summarization. We also block repeated 4-grams in summarization outputs.

5 Development Results

We report exact match (EM), chrF (Popović, 2015), ROUGE-L (Lin, 2004), and LaBSE cosine similarity (Feng et al., 2022). For open QA, we use chrF, ROUGE-L, and LaBSE together with manual inspection of validation outputs.

5.1 Task-Level Results

Table 2 reports development scores for the evaluated adapters. The 8.5k context mixture and the 12k summary mixture are selected for submission.

The targeted context continuation is marginally higher on ROUGE-L and LaBSE, whereas the 8.5k context mixture is higher on EM and chrF. We therefore select the latter.

5.2 Scientific Summarization

Table 3 reports results on a 336-example scientific validation set. Gold targets are author abstracts; non-English targets are translated and quality filtered. The 12k mixture is strongest across all three metrics.

Blocking repeated 4-grams removes the observed summary loops.

5.3 Open QA and Submitted Systems

Best-score QA has the strongest aggregate automatic scores. Long-form QA is stronger on a manu-

Task	Training variant	n	EM	chrF	ROUGE-L	LaBSE
Summarization	Multitask adapter	1,651	0.00	22.97	0.1571	0.6681
	12k general/scientific mix	1,651	0.00	**26.81**	**0.1780**	**0.6919**
Context QA	Multitask adapter	1,517	66.25	77.07	0.6834	0.8813
	8.5k context mixture	1,517	**68.89**	**78.58**	0.6959	0.8897
	+4.8k targeted continuation	1,517	68.36	78.36	**0.6966**	**0.8906**
Open QA	Multitask adapter	794	3.90	28.34	0.2287	0.6524
	Long-form QA	794	4.03	27.96	0.2307	0.6600
	Best-score QA	794	**4.16**	**28.41**	**0.2329**	**0.6638**

Table 2: Internal model-selection results on identical held-out rows. EM is a percentage and LaBSE is cosine similarity.

Model	chrF	ROUGE-L	LaBSE
Multitask adapter	13.27	0.0742	0.6556
8k summary mix	25.28	0.1812	0.6959
12k summary mix	**29.76**	**0.2082**	**0.7535**

Table 3: Results on the scientific summarization validation set ($n = 336$). The summary mixtures are evenly divided between general and scientific data.

System	Summary	Context	Open
PSK-primary	12k mix	8.5k mix	Long-form
PSK-variant_A	12k mix	8.5k mix	Best-score
PSK-variant_B	12k mix	8.5k mix	Multitask

Table 4: Submitted routes. The open-QA variants are named for their selection criterion.

ally identified set of long-form prompts and reaches the generation limit less often. Our main submission uses Long-form QA. A second submission uses Best-score QA. Table 4 summarizes the three submitted routes.

6 Analysis

Separate adapters improve context QA and summarization on our development set. Scientific summarization also improves after adding papers and their abstracts to the training data. Open QA has no clear winner. Best-score QA performs better on automatic metrics, while Long-form QA works better on the longer questions we checked manually and is less likely to hit the output limit.

The shared summarization route improves chrF over the multitask adapter in all 28 represented languages. The shared context-QA route improves exact match in 19 of 25 languages, ties in four, and declines in two. Open QA remains mixed: the primary route improves chrF in 9 of 19 languages, and variant A improves it in 10 of 19. Appendix A gives the complete language-level changes and cell sizes.

7 Conclusion

We present a routed Tiny Aya Global system with one QLoRA adapter per MIST task. Development results select the 8.5k context mixture and the 12k summary mixture, while the three submissions vary the less stable open-QA route. Official evaluation will provide the appropriate shared-task comparison.

Limitations

Reference metrics do not directly measure factuality and are particularly limited for open QA. Translated scientific targets may retain artifacts. Results use one backbone and one development split. The router also assumes that the task label is available.

Ethics Statement

Public and challenge-provided datasets are used under their respective terms. Machine translation may reproduce source and model biases, and open-ended outputs may contain incorrect or harmful claims. The system should not be treated as a factual authority without verification.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat

Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Pranjal A. Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. 2026. UPDESH: Synthesizing grounded instruction tuning data for 13 Indic languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37997–38041, San Diego, California, United States. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115, New Orleans, Louisiana, USA. Neural Information Processing Systems Foundation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2026. MCIF: Multimodal crosslingual instruction-following benchmark from scientific talks. In *The Fourteenth International Conference on Learning Representations*.

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.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. Tiny Aya: Bridging scale and multilingual depth. *Preprint*, arXiv:2603.11510.

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

WMT 2026 MIST Organizers. 2026. Multilingual instruction shared task. `https://www2.statmt.org/wmt26/multilingual-instruction.html`. Accessed 7 August 2026.

A Language-Level Changes

Tables 5 and 6 report changes from the multitask adapter on identical held-out examples. All submitted systems share the summary and context routes. The primary system uses Long-form QA, variant A uses Best-score QA, and variant B retains the multitask open-QA route, for which every open-QA change is zero. Context-QA changes are exact-match percentage points; summary and open-QA changes are chrF points. These are descriptive results, and some language cells are small.

Language	n_S	Δ_S	n_C	Δ_C
Arabic	85	+4.39	90	+4.44
Bengali	66	+8.18	20	+10.00
Bhojpuri	14	+26.89	–	–
Czech	14	+12.80	–	–
Central Kurdish	14	+9.23	45	-2.22
German	34	+5.97	93	+3.23
English	70	+0.87	93	+3.23
Finnish	14	+6.59	45	+2.22
French	68	+1.11	45	+0.00
Haitian Creole	34	+1.86	45	+2.22
Hindi	85	+5.40	45	+2.22
Indonesian	85	+3.26	90	+4.44
Italian	34	+3.83	93	+2.15
Japanese	82	+3.02	90	+4.44
Korean	79	+2.03	90	-1.11
Marathi	79	+5.53	45	+0.00
Persian	69	+5.59	45	+6.67
Portuguese	84	+0.09	45	+2.22
Russian	85	+5.10	90	+2.22
Slovak	34	+8.72	45	+2.22
Spanish	85	+4.17	45	+4.44
Swahili	53	+1.88	45	+4.44
Telugu	53	+0.82	45	+0.00
Thai	48	+2.30	45	+2.22
Turkish	84	+2.45	45	+8.89
Vietnamese	84	+4.07	45	+2.22
Yoruba	47	+2.60	–	–
Chinese	68	+1.93	93	+0.00

Table 5: Changes for the shared routes. S denotes summarization (ΔchrF), and C denotes context QA (ΔEM in percentage points).

Language	n	**Primary ΔchrF**	**Var. A ΔchrF**
Arabic	52	+0.22	-0.16
Bengali	11	+0.15	-0.72
Czech	11	+1.53	+1.62
German	43	-2.02	-1.34
English	11	-0.18	+1.82
French	45	-1.89	-0.79
Hindi	52	-1.44	+1.75
Indonesian	53	+0.86	+1.50
Italian	45	+2.94	+2.16
Japanese	52	-1.25	-0.84
Korean	45	-0.43	+0.02
Marathi	45	-4.24	-2.62
Portuguese	45	-0.70	-0.68
Russian	52	-1.65	-1.11
Spanish	45	-2.42	-1.57
Turkish	45	+0.21	+1.33
Vietnamese	45	+1.62	+1.00
Yoruba	45	+1.87	+1.57
Chinese	52	+1.38	+0.03

Table 6: Open-QA chrF changes. The primary system uses Long-form QA; variant A uses Best-score QA. Variant B is the multitask reference and has zero change.