

TUM-Heilbronn: WMT26 Multilingual Instruction Shared Task

Jen-Chien Hou¹ and Tsedeniya Kinfe Temesgen^{1,2} and Alexander Fraser^{1,2}

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

²Munich Center for Machine Learning (MCML)
{jenchien.hou|tsedeniya.temesgen|alexander.fraser}@tum.de

Abstract

This paper describes the TUM-Heilbronn system submission to the WMT26 Multilingual Instruction Shared Task (MIST). The MIST covers three tasks: context-based question answering, open-ended generation, and cross-lingual summarization to evaluate the multilingual instruction-following ability of large language models. We fine-tune Qwen3.5-9B-Instruct on 27 diverse languages, across 10 different writing scripts. Our model outperformed 6 of the 15 models submitted to this shared task. Moreover, we ranked first in truthfulness (i.e., decline to answer when the context lacks sufficient information) on the context-based question-answering task. Cross-lingual summarization remains a challenging task for all submitted models, with the lowest performance observed among the three tasks.

1 Introduction

Large Language Models (LLMs) have demonstrated a remarkable capability in following instructions and performing a wide range of tasks effectively. However, this capability is closely associated with the language in which the instructions are provided. Models generally exhibit stronger instruction-following performance in high-resource languages, particularly English (Dey et al., 2024), whereas their performance tends to decline for mid- and low-resource languages (Zeng et al., 2025). This disparity is even more pronounced for low-resource languages that use non-Latin writing systems.

Recent studies show that instruction-following abilities do not transfer equally across languages. Zeng et al. (2025) find a performance gap between high- and low-resource languages, with model size and script also influencing language adherence. Similarly, Le et al. (2025) show that multilingual instruction following becomes particularly challenging for complex information extraction tasks,

QA-Context	Answer the following question concisely given the English content: [Context and question are provided here] Answer: [answer in English]
OEG	Tell me five fun facts about Grand Canyon?, Answer: [answer in English] Absolutely! Here are five fun facts about the Grand Canyon: **It's Really Old:**
SUM	Fasse den englischen Inhalt in einem Abstract mit ungefähr 200 Wörtern zusammen. (Summarize the English content in an abstract of approximately 200 words.) [Content in English] Answer: [answer in German]

Table 1: Examples of multilingual instruction following across *three* tasks: QA-Context (context-based question answering), OEG (open-ended generation), and SUM (summarization). English translations of the text are provided in parentheses.

while Liu et al. (2025), Dussolle et al. (2025), and Dong et al. (2026) report further disparities across languages. Closing these performance gaps is essential for developing LLMs that can reliably follow instructions across languages without degrading performance on NLP downstream tasks, ultimately enabling users to interact with models in their own language.

The WMT26 Multilingual Instruction Shared Task (MIST) (Schmidtová et al., 2026) aims to address this challenge by evaluating multilingual instruction-following capabilities of LLMs across diverse languages, scripts, and tasks. MIST proposes three tasks: **Context-based Question Answering** (QA-Context), **Open-Ended Generation** (OEG), and **Cross-lingual Summarization** (SUM). The first two tasks evaluate the model’s ability to provide accurate and faithful responses in the input language, while the third evaluates cross-

lingual generation, where the model is expected to give a summary in a language different from the input document, as illustrated in Table 1.

The shared task covers 27 high- and low-resource languages. High-resource languages include Czech, German, English, Finnish, French, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Portuguese, Russian, Slovak, Spanish, Vietnamese, and Chinese (Simplified). Low-resource languages Arabic, Bengali, Central Kurdish, Haitian Creole, Marathi, Swahili, Telugu, Thai, Turkish, and Yoruba are covered. A surprise language, Bhojpuri, was introduced to assess models’ ability to generalize to unseen linguistic contexts. In addition, submissions are restricted to models up to 10B parameters.

Our submission investigates the multilingual instruction-following capabilities of **Qwen3.5-9B**¹ model, across three tasks. The model is jointly fine-tuned on all tasks using the parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2021) technique. Our results show stronger performance on QA-Context and OEG compared to SUM. On QA-Context, our model achieves the highest BLEU and ROUGE-L scores among all submitted models². It also performs best on truthfulness, which measures models’ ability to decline to answer when the context lacks sufficient information. On OEG, our model performs relatively well at providing answers in the target language (i.e., language adherence), although its overall task performance is low. SUM remains the most challenging task, with all submitted models performing worse compared to the other two tasks. In addition, the low performance on the surprise language, Bhojpuri, indicates that the model still struggles to generalize to unseen languages.

2 Tasks and Languages

MIST comprises three tasks covering 27 diverse languages to evaluate instruction-following abilities of LLMs. For training, the organizers provide a dataset³, which we extend with additional data discussed in the following sections.

¹<https://huggingface.co/Qwen/Qwen3.5-9B>

²We compare our results against the best-performing models submitted to this shared task, as gold labels for the test set are not available to us.

³<https://huggingface.co/datasets/pinzhenchen/wmt26-mist-sample>

Task	Source	Train	Val	Test
QA-Context	Belebele TyDI QA MCIF	7,545	1,915	8,640
OEG	Aya WMT25-OEG	4,129	1,034	2,359
SUM	CrossSum WikiLingua MCIF	8,122	904	1,776
Total		20,589	3,940	12,775

Table 2: Dataset source and statistics used across tasks.

2.1 Context-based Question Answering (QA-Context)

In this task, the model is prompted to answer a question by inferring from a given context, as shown in the example in Table 1. If the context is insufficient to answer reliably, the model should refrain from responding (i.e., truthfulness).

Dataset: The training dataset provided by the shared task organizers is a subset of the original data sourced from Belebele (Bandarkar et al., 2024), TyDi QA (Clark et al., 2020), and MCIF (Papi et al., 2026), a multimodal cross-lingual instruction-following dataset covering recognition, translation, question answering, and summarization subtasks, from which only the question-answering subtask was selected.

To extend this dataset, we include additional samples from each source. We further augment the dataset to capture *truthfulness* by introducing a small subset of unanswerable instances, created by replacing the original passages with passages drawn from different articles. Across all three sources, we select an equal number of instances per language and prepare a training and validation set by splitting the data in an 80:20 ratio.

2.2 Open-ended Generation (OEG)

Given an open-ended question, the model is expected to provide a natural, helpful response in the same language as the question, as illustrated by the example in Table 1.

Dataset: The datasets are collected from Aya dataset (Singh et al., 2024) and the WMT25 Multilingual Instruction Shared Task, hereafter WMT25-OEG (Kocmi et al., 2025).

We extend this data from the same source by retaining instances where the instruction and answer are in the same language, as OEG is designed to

be a monolingual task where the instruction and response are expected to be in the same language. Finally, we use an 80:20 ratio to split the collected data into training and validation sets.

2.3 Cross-lingual Summarization (SUM)

This is a cross-lingual task in which the model is instructed, in one language (e.g., German), to summarize a document written in another language (e.g., English), as shown in the example in Table 1. The training data also includes examples in which the output language differs from both the instruction and context languages. For example, an English instruction, "*Summarize the text in Spanish*", is paired with a Chinese context.

Dataset: We used the dataset provided by the organizers, which was collected from Cross-Sum (Hasan et al., 2021), WikiLingua (Ladhak et al., 2020), and the summarization part of the MCIF dataset (Papi et al., 2026), and we split it into training and validation using a 90:10 ratio.

Overall, we collected 24K examples: 20K for training and 3.9K for validation. (Cf. Table 2). The test set, provided by the organizers, contains $\approx 13K$ instances, with instructions given in 24 languages, including a surprise language, Bhojpuri.

2.4 Languages

In total, the training dataset covers 27 languages. Of these, 17 languages, namely Arabic, Bengali, Chinese, English, German, Hindi, Indonesian, Italian, Japanese, Korean, Marathi, Portuguese, Russian, Spanish, Turkish, Vietnamese, and Yoruba, are covered across all three tasks. Four languages, Finnish, Swahili, Telugu, and Thai are each covered by only one task, while the remaining six, Central Kurdish, Czech, French, Haitian Creole, Persian, and Slovak, are covered by exactly two tasks. (Cf. Table 12 for list of languages per task)

Overall, 23 languages are covered in QA-Context and SUM, while 17 languages are included in the OEG task for training. The test set covers 24 languages, excluding French, Swahili, Telugu, and Thai, and introduces Bhojpuri as a surprise language to evaluate the model’s ability to generalize to unseen languages during training.

3 System Description

Experimental setup: We fine-tune the Qwen3.5-9B⁴ model on three tasks. We use a 4-bit-quantized Qwen3.5-9B model and apply Low-Rank Adaptation (LoRA) with a rank of $r = 16$ on a single NVIDIA A100 GPU. Training data were reformatted as chat-style instruction pairs. We first fine-tuned for 2 epochs on QA-Context and OEG, then fine-tuned for an additional 3 epochs on the SUM dataset. Loss was computed only on the assistant responses. (Cf. Table 7 for hyperparameters).

Evaluation metrics: According to the shared-task scoring scheme, the Ensemble and Gated ROUGE-L scores are used as primary evaluation metrics. In addition, BLEU, ROUGE-L, and chrF++ scores are reported for comparison.

- **Ensemble:** For QA-Context, verification accuracy is determined by a majority vote among three LLM judges: *GLM-1*, *Command-A+*, and *Gemma-4-31B*. OEG task uses these three judges together with *Qwen3.5-397B*. A prediction is considered correct if at least two judges classify it as correct. (For detail refer Schmidová et al. (2026))
- **Gated ROUGE-L:** This metric first applies language identification GlotLID (Kargaran et al., 2023) to evaluate whether the model response is in the expected language, then computes ROUGE-L. If the response is not in the expected target language, the instance is assigned a score of 0.0, regardless of its ROUGE-L scores. This metric is used to evaluate the SUM task.
- **Lexical overlap:** BLEU (Papineni et al., 2002), chrF++ (Popović, 2015) and ROUGE-L (Lin, 2004) is computed for the QA-Context and SUM task.

4 Preliminary Experiment

We report results on the validation set using BERTScore (Zhang et al., 2020). Our fine-tuned model improves BERTScore by 12% and 6.3% on the QA-context and SUM tasks, respectively. On OEG, it achieves 72.87%, a 7.34% gain over the baseline (Table 9). These improvements provide empirical support to adopt the above experimental

⁴<https://huggingface.co/Qwen/Qwen3.5-9B>

	Overall Rank	QA-Context	Rank	OEG	Rank	SUM	Rank
CUNI-variant_B	1	80.93	2	88.45	1	17.41	2
CUNI-primary	2	81.63	1	82.25	2	14.71	5
Lanigo-primary	3	76.90	3	76.27	3	16.95	3
Lanigo-variant_A	4	69.80	4	64.26	7	17.55	1
Sentinel-base	5	59.58	7	72.47	4	14.11	7
PSK-primary	6	51.70	10	63.65	8	15.66	4
Sentinel-primary	7	50.84	11	65.43	5	14.11	6
Sentinel-RunB	8	50.84	12	65.43	6	13.44	8
TUM-Heilbronn-primary	9	59.16	8	60.48	10	12.50	9
TeamSpinoza-Variant_V4	10	61.96	6	57.79	11	10.87	12
Aleix-variant_B	11	57.08	9	62.37	9	7.81	14
TeamSpinoza-primary	12	67.50	5	47.91	14	7.42	15
Aleix-variant_A	13	45.73	13	53.66	13	12.06	11
Aleix-primary	14	34.73	14	55.73	12	8.85	13
TeamSpinoza-Variant_sum-verifier3	15	17.22	15	23.55	15	12.33	10

Table 3: **Leaderboard.** Ranking of all 15 systems across the 3 tasks. **Overall Rank** denotes the mean rank across all tasks, while each task block reports the model’s performance and corresponding rank for that task. For QA-Context and OEG, performance is reported using the ensemble (LLM-judge majority vote). For SUM, Gated ROUGE-L is reported. Our system is highlighted in gray and the best scores are in bold.

setup for the test-set evaluation. Section 5 discusses results on the test set.

5 Experimental Results

We compare our model with the best-performing models submitted to this shared task on the *test set* across all three tasks⁵. Overall, *CUNI-variant_B* achieves the best performance on average and ranked first on the OEG task. *CUNI-primary* performs best on the QA-Context task, while *Lanigo-variant_A* achieves the strongest performance on the SUM task. (Cf. Leaderboard in Table 3).

Our model achieves the lowest performance on the SUM task, with a Gated ROUGE-L score of 12%. In contrast, it achieves substantially higher scores on the QA-Context and OEG tasks, with scores of 59.16% and 60.48%, respectively.

QA-Context: Although our system ranks eighth in the ensemble score, it obtains the highest BLEU and ROUGE-L scores as shown in Table 4. Moreover, our model ranks first on truthfulness-F1, suggesting that it is less likely to generate unsupported answers or hallucinations, as shown in Table 5.

OEG: The system achieves an ensemble score of 60.48%, while providing answers in the expected target language in 87% of cases, shown in Table 8. Despite prior findings that models tend to fall back to English when instructed in non-English lan-

	BLEU	chrF++	ROUGE-L (LCS F1)
CUNI-primary	8.7	35.1	24.14
CUNI-variant_B	10.6	38.2	26.96
Lanigo-primary	9.5	36.1	24.88
Lanigo-variant_A	10.4	34.9	25.14
TUM-Heilbronn-primary	19.7	34.0	34.95

Table 4: **QA-Context:** Performance of the TEST SET measured using BLEU, chrF++, and ROUGE-L scores. and ROUGE-L scores. Best scores are in bold.

	Rank	F1 (%)
TUM-Heilbronn-primary	1	87.07
TeamSpinoza-Variant_V4	2	86.50
TeamSpinoza-primary	3	86.40
CUNI-primary	4	82.02
CUNI-variant_B	5	81.79
Lanigo-primary	6	80.41
Lanigo-variant_A	7	76.72
Sentinel-primary	8	64.72
Sentinel-RunB	9	64.72
PSK-primary	10	63.40

Table 5: **Truthfulness:** F1 score on unanswerable questions.

guages (Marchisio et al., 2024), our results demonstrate a strong ability in instruction following and language understanding.

SUM: Our model achieves a Gated ROUGE-L score of 12.5%, and performs considerably worse on lexical-overlap metrics (See Table 6). In addition, our model achieves 75.3% accuracy in providing answers in the target language. (See Table 8). Overall, the results indicate significant challenges in the SUM task in generating suitable summaries

⁵We omit baseline results on the test set as gold labels are not available to us. Instead, we compare our results with other participants of this shared task. In addition, we report BERTScore on our validation data .

and, to a lesser extent, in consistently responding in the target language.

For the surprise language, Bhojpuri, our model achieves scores of 52.2% on QA-Context, 29.6% on OEG, and 9.5% on SUM. These results are substantially below the best-performing models, indicating a limited ability to generalize to unseen languages. (Cf. Table 11, for per-language performance and Table 10 for each LLM judge scores)

	BLEU	chrF++	ROUGE-L
Lanigo-variant_A	6.5	34.5	19.23
CUNI-variant_B	6.5	38.3	18.93
Lanigo-primary	6.4	37.8	18.51
PSK-primary	4.2	28.1	17.30
TUM-Heilbronn-primary	4.4	25.8	14.58

Table 6: **SUM**: Performance of the TEST SET measured using BLEU, chrF++, and ROUGE-L scores. and ROUGE-L scores. Best scores are in **bold**.

6 Conclusion

We presented our submission to the WMT26 Multilingual Instruction Shared Task, in which we participated in three tasks: Context-based Question Answering (QA-Context), Open-Ended Generation (OEG), and Summarization (SUM) covering diverse languages and scripts. We fine-tuned the Qwen3.5-9B model on the three tasks and evaluated its instruction-following ability across 23 languages and one surprise language. The results show that our model performs differently across tasks, with stronger performance on QA-Context and OEG than on SUM. In the QA context, the model achieves competitive results and performs best on the truthfulness F1 score. In OEG, it demonstrates relatively good language adherence despite lower overall task performance. In contrast, SUM remains challenging, with a substantially lower overall ensemble (Majority vote) score. Performance also drops considerably for the surprise language, Bhojpuri, across all three tasks. Overall, these results highlight the challenges of multilingual instruction following, particularly for less-resourced languages and cross-lingual generation tasks.

7 Limitations

Model: We evaluate only a single model, i.e., Qwen3.5-9B, across the three tasks. It is important to evaluate smaller and different architecture models, as performance and cross-lingual generalization may vary substantially. We note that com-

paring multiple model families is relevant to assessing whether our findings generalize beyond this specific model.

Training strategy: We do not experiment with hyperparameter tuning. All hyperparameters. A more systematic hyperparameter search could improve performance.

Analysis of relation between tasks: We do not examine the interactions between tasks, such as whether one task complements or comes at the cost of another. However, we observe one such interaction. Further training on summarization data leads to shorter open-ended answers that are less likely to meet the requested sentence length, likely due to the summarization objective biasing the model toward concise outputs.

Understanding these interactions could shed light on shared structure across tasks, but we leave this investigation for future work.

Analysis on the relation between languages: We did not analyze how languages influence one another in instruction-following tasks. Since the model is fine-tuned jointly on all languages, the presence of certain languages may affect performance on others through shared representations, and this interaction could be shaped by factors such as script, language family, or resource availability. Understanding these effects could help explain performance disparities across languages.

8 Acknowledgments

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

- 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.
- 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.
- Krishno Dey, Prerona Tarannum, Md. Arif Hasan, Imran Razzak, and Usman Naseem. 2024. [Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings](#). *ArXiv*, abs/2410.13153.
- Jianshuo Dong, Yutong Zhang, Liu Yan, Zhenyu Zhong, Tao Wei, Chao Zhang, and Han Qiu. 2026. [Revisiting the reliability of language models in instruction-following](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7784–7812, San Diego, California, United States. Association for Computational Linguistics.
- Antoine Dussolles, Andrea Cardeña Díaz, Shota Sato, and Peter Devine. 2025. [M-IFEval: Multilingual instruction-following evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6176–6191, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong bin Kang, and Rifat Shahriyar. 2021. [Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs](#). *CoRR*, abs/2112.08804.
- Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. [Evaluating the elementary multilingual capabilities of large language models with MultiQ](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- 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 (WMT 2025)*, pages 462–483, 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.
- Thang Le, Huy Huu Nguyen, Anh Tuan Luu, and Thien Huu Nguyen. 2025. [Massively multilingual instruction-following information extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3542–3585, Vienna, Austria. 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.
- Yile Liu, Ziwei Ma, Xiu Jiang, Jinglu Hu, ChangJing ChangJing, and Liang Li. 2025. Maxife: Multilingual and cross-lingual instruction following evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14252–14332.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. [Understanding and mitigating language confusion in LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.
- Sara Papi, Maïke 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*.
- 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.

Patrícia Schmidtová, Katia Artemova, Seth Ayccock, Niyati Bafna, Tom Kocmi, Philipp Koehn, Danni Liu, Nam Luu, Sara Papi, Beatrice Savoldi, Mariya Shmatova, Vilém Zouhar, Maike Züfle, and Pinzhen Chen. 2026. Findings of the WMT26 multilingual instruction shared task: Small models are not yet polyglots. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.

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 Het-tiarachchi, 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.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. [Marco-bench-MIF: On multilingual instruction-following capability of large language](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.

A Hyperparameters

Table 7 summarizes the hyperparameters that were used to fine-tune the Qwen3.5-9B model.

Hyperparameter	Value
Learning rate	2×10^{-4}
LoRA r	16
Alpha α	32
Batch size	16
Epochs	$\dagger 5$
Optimizer	AdamW
Seed	42

Table 7: Hyperparameters used for fine-tuning. $\dagger$ The model was initially fine-tuned for 2 epochs on QA-Context + OEG and continued fine-tuning for 3 epochs SUM, i.e., a total of 5 epochs.

B Truthfulness

In Context-based Question Answering, truthfulness measures the model’s ability to refrain from answering when the given context does not contain enough information to answer the question. We report the results in Table 5.

C Language Identification

Previous studies have shown that multilingual models tend to fall back to English when prompted in less well-supported languages (Marchisio et al., 2024; Holtermann et al., 2024). To assess whether the generated responses are in the expected target language, GlotID (Kargaran et al., 2023) is used. Results are reported in Table 8.

Task	Accuracy
QA-Context	92.2
OEG	87.7
SUM	75.3

Table 8: Language identification (GlotID) accuracy

D LLM Judge Accuracy

Table 10 shows the scores by each judge across tasks.

E Preliminary result on Validation set

	Baseline		Finetuned	
	BERTScore	ROUGE-L	BERTScore	ROUGE-L
QA-Context	69.97	15.17	82.57	28.21
SUM	66.55	9.54	72.87	15.78
OEG	65.53	-	71.08	-

Table 9: Results on the dev set across tasks. ROUGE-L is reported for comparison for QA-Context and OEG tasks.

	Ensemble [Majority Vote]	Command-A+ [Judge]	GLM-1 [Judge]	Gemma-4-31B [Judge]	Qwen 3.5 397B [Judge]
QA	59.16	61.59	54.48	64.05	–
OEG	60.48	58.26	61.28	60.32	63.57
SUM	–	52.36	47.41	–	–

Table 10: LLM judge verification accuracy (%) across tasks. For SUM, GLM-1, and Command-A+, judges evaluate target-language fluency and grammar

F Performance of each language

The performance of our model vs. the top 3 models submitted to this shared task. The results reported in Table 11 covering three tasks across 25 languages.

	QA-Context				OEG				SUM			
	CUNI- primary	CUNI- variant_B	Laniquo- primary	TUM- Heilbronn- primary	CUNI- variant_B	CUNI- primary	Laniquo- primary	TUM- Heilbronn- primary	Laniquo- variant_A	CUNI- variant_B	Laniquo- primary	TUM- Heilbronn- primary
Arabic	84.70	84.70	81.70	60.30	92.40	87.90	80.30	65.30	16.00	14.60	14.70	13.90
Bengali	84.40	85.60	82.20	55.30	93.30	79.30	82.50	54.50	14.80	14.90	14.70	8.60
[†] Bhojpuri	81.10	85.80	80.00	52.20	88.90	59.20	80.30	29.60	17.30	18.00	18.60	9.50
Chinese	84.20	83.60	75.80	63.10	90.10	93.30	85.70	78.00	27.90	25.90	26.10	25.70
Czech	83.10	79.70	78.30	66.40	86.00	78.00	65.90	65.30	14.90	13.70	14.30	8.30
English	86.90	86.40	82.80	76.40	93.60	96.80	85.70	82.20	24.80	22.00	22.30	25.60
Finnish	83.90	77.50	74.70	58.60	88.50	70.10	70.70	52.90	12.10	12.20	11.70	4.50
German	85.30	86.90	81.90	66.40	93.90	91.70	86.00	71.00	17.10	15.00	16.10	3.90
Haitian	80.00	71.70	67.80	51.90	86.60	66.20	67.20	39.50	20.20	19.90	19.00	16.00
Hindi	83.60	85.30	83.10	60.80	93.30	81.50	83.10	56.40	21.90	21.10	20.00	14.80
Indonesian	85.80	85.00	82.50	60.30	92.70	93.60	83.10	75.80	22.40	20.90	21.40	9.70
Italian	86.70	88.10	82.50	70.80	92.70	93.60	83.10	77.40	21.70	19.70	20.00	19.70
Japanese	84.40	82.50	78.60	61.70	88.90	89.80	83.40	62.10	2.30	6.30	3.10	5.90
Korean	83.90	83.60	81.70	60.00	88.20	82.50	78.00	62.40	15.40	14.80	14.50	10.80
Kurdish	71.10	73.60	66.70	32.20	78.70	56.70	51.90	24.50	11.40	14.00	13.30	5.50
Marathi	76.40	86.70	80.60	51.90	92.70	64.60	81.20	41.40	12.90	13.30	12.50	7.80
Persian	85.00	82.80	78.90	63.30	91.70	93.60	82.80	64.30	20.50	19.20	19.80	15.90
Portuguese	85.80	86.10	81.40	67.80	91.40	94.60	83.10	72.60	23.80	22.40	21.50	15.20
Russian	83.60	83.30	78.10	65.80	93.00	95.20	82.20	76.80	16.00	15.50	15.10	13.20
Slovak	83.10	79.70	73.60	61.70	72.60	75.20	52.20	56.40	14.60	13.50	14.50	9.00
Spanish	85.60	85.30	82.20	68.90	92.00	94.90	85.40	77.40	26.20	25.50	25.10	18.40
Turkish	83.30	84.70	80.30	58.90	92.00	84.70	76.10	60.80	16.80	17.10	15.40	11.00
Vietnamese	84.20	85.00	85.60	65.00	88.50	91.10	78.00	71.30	23.60	23.40	21.90	21.80
Yoruba	43.10	28.60	24.70	20.00	41.20	43.40	18.10	14.30	6.70	14.80	11.40	5.50
Overall	81.60	80.90	76.90	59.20	88.50	82.30	76.30	60.50	17.60	17.40	17.00	12.50

Table 11: Performance across languages and the three tasks, comparing our results with those of the three best-performing systems submitted to the shared task. The surprise language denoted by [†], a language not seen during fine-tuning

Language	Code	QA-Context			SUM			OEG			All
		Train	Val	Test	Train	Val	Test	Train	Val	Test	
Arabic	arb_Arab	474	99	360	428	46	74	283	63	100	1927
Bengali	ben_Beng	458	120	360	318	29	74	37	9	100	1505
Central Kurdish	ckb_Arab	223	59	360	—	—	74	—	—	100	816
Chinese (Simpl.)	zho_Hans	405	102	360	409	50	74	288	58	100	1846
Czech	ces_Latn	217	51	360	—	—	74	37	9	100	848
English	eng_Latn	614	163	360	432	44	74	37	9	100	1833
Finnish	fin_Latn	419	129	360	—	—	74	—	—	100	1082
[†] French	fra_Latn	—	—	—	412	39	—	230	70	—	751
German	deu_Latn	424	95	360	178	22	74	227	60	100	1540
Haitian Creole	hat_Latn	207	55	360	85	15	74	—	—	100	896
Hindi	hin_Deva	217	53	360	411	61	74	249	61	100	1586
Indonesian	ind_Latn	438	125	360	434	39	74	279	67	100	1916
Italian	ita_Latn	397	89	360	174	26	74	226	74	100	1520
Japanese	jpn_Jpan	435	107	360	412	41	74	282	64	100	1875
Korean	kor_Hang	479	116	360	385	48	74	238	61	100	1861
Marathi	mar_Deva	222	53	360	391	40	74	230	69	100	1539
Persian	pes_Arab	207	64	360	333	36	74	—	—	100	1174
Portuguese	por_Latn	205	61	360	424	45	74	247	53	100	1569
Russian	rus_Cyrl	443	104	360	424	47	74	278	68	100	1898
Slovak	slk_Latn	215	47	360	88	12	74	—	—	100	896
Spanish	spa_Latn	230	52	360	428	42	74	240	60	100	1586
[†] Swahili	swh_Latn	—	—	—	323	32	—	—	—	—	355
[†] Telugu	tel_Telu	—	—	—	320	34	—	—	—	—	354
[†] Thai	tha_Thai	—	—	—	291	31	—	—	—	—	322
Turkish	tur_Latn	206	61	360	411	39	74	241	59	100	1567
Vietnamese	vie_Latn	208	55	360	416	48	74	237	63	100	1561
Yoruba	yor_Latn	202	55	360	195	22	74	243	57	59	1267
[‡] Bhojpuri	bho_Deva	—	—	360	—	—	74	—	—	100	534
Total		7545	1915	9000	8122	904	1850	4129	1034	2459	36958

Table 12: Languages covered by the WMT26 Multilingual Instruction Shared Task, with ISO 639-3 + script codes and number of instances per language. [†] indicates languages not included in the test set, and [‡]Bhojpuri is a surprise language only covered in the test set. Light green indicates languages covered by only one task and light pink indicate languages covered by two task. The uncolored languages are covered by all three tasks. Overall, 23 languages are covered by QA-Context and SUM. 17 languages are in OEG.