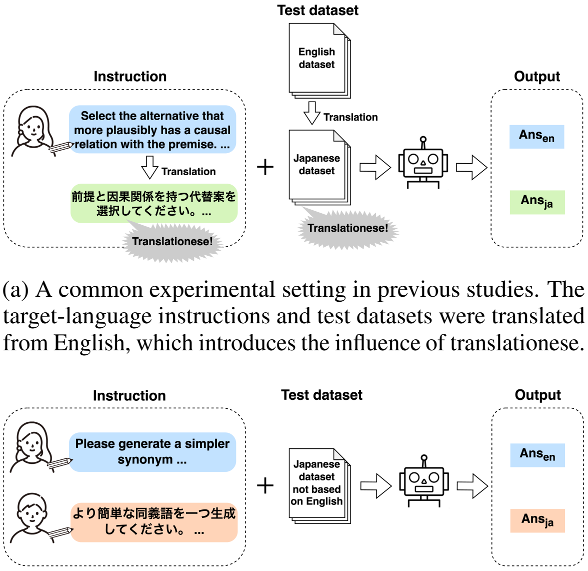
A Fair Comparison without Translationese: English vs. Target-language Instructions for Multilingual LLMs

Taisei Enomoto[1], Hwichan Kim[1], Zhousi Chen[2], Mamoru Komachi[2]
[1]Tokyo Metropolitan University [2]Hitotsubashi University
{enomoto-taisei, kim-hwichan}@ed.tmu.ac.jp
{zhousi.chen, mamoru.komachi}@r.hit-u.ac.jp

Abstract

Most large language models are multilingual instruction executors. Prior studies suggested that English instructions are more effective than target-language instructions even for non-English tasks; however, these studies often use datasets and instructions translated from English, which introduce biases known as *translationese*, hindering an unbiased comparison. To address this issue, we conduct a fair comparison between English and target-language instructions by eliminating translationese effects. Contrary to previous studies, our experiments across several tasks reveal that the advantage of adopting English instructions is not overwhelming. Additionally, we report on the features of generated texts and the instruction-following abilities when using respective instructions. Our source code is publicly available at the following URL[1].


Instruction
Select the alternative that more plausibly has a causal relation with the premise. ...
Translation
前提と因果関係を持つ代替案を選択してください。...
Translationese!
+
Test dataset
English dataset
Translation
Japanese dataset
Translationese!
Output
Ans_en
Ans_ja
Instruction
Please generate a simpler synonym ...
より簡単な同義語を一つ生成してください。 ...
+
Test dataset
Japanese dataset not based on English
Output
Ans_en
Ans_ja

(a) A common experimental setting in previous studies. The target-language instructions and test datasets were translated from English, which introduces the influence of translationese.

(b) Experimental setting of this study. The fair instruction construction process is described in Section 3.1.

Figure 1: Overview of experiments from (a) previous studies and (b) this study.

1 Introduction

In recent years, large language models (LLM) have demonstrated outstanding performance across a variety of natural language processing (NLP) tasks. To fully leverage their capabilities, it is crucial to provide these models with appropriate instructions (Wang et al., 2024; Niwa and Iso, 2024). Specifically, because multilingual LLMs (MLLM) offer better non-English performance, an unavoidable question—*should instructions be given in English or the target-language?*— has been under discussion in several studies (Lin et al., 2022; Muennighoff et al., 2023; Ahuja et al., 2023). A reasonable consideration of this issue is that the training process for MLLMs is still dominated by English data, suggesting that English instructions might be more effective, even for non-English tasks. Indeed, previous studies have reported the effectiveness of English instructions by comparing the languages used in instructions for MLLMs (Muennighoff et al., 2023; Ahuja et al., 2023).

However, a flaw exists in these studies: the target-language datasets and instructions were produced by translating from English (Figure 1a). Texts produced through translation are prone to information loss, unnatural expressions, and stylistic differences compared to texts written by native speakers—phenomena referred to as "*translationese*" [2] (Lembersky et al., 2012; Eetemadi and Toutanova, 2014; Wintner, 2016; Clark et al., 2020). Consequently, target-language datasets translated from English may exhibit expressions that resemble English writing style or contain content influenced by the cultural and contextual background of English-speaking countries. Additionally, target-language instructions translated from English may contain different information. These factors in-

[1] https://github.com/enomooon/fair_comparison_instructions

[2] Appendix A shows examples of translationese.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 649–670
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

dicate the possibility that English instructions in previous studies were inherently advantaged, making it likely that comparisons between English and target-language instructions were biased.

To this end, our study aims to conduct a fair comparison between English and target-language instructions in MLLMs, by eliminating translationese effects. Specifically, we leverage target-language datasets and instructions that are not translated from English to investigate performance differences across a range of tasks (Figure 1b). In particular, for the classification task, we employ multiple classification label sets to explore changes resulting from variations in the label sets. Our experimental results reveal that, contrary to previous studies, whether English or target-language instructions tend to perform better depends on the task and labels. Additionally, we conduct a detailed analysis comparing the features of generated texts and the instruction-following abilities of MLLMs when using English versus target-language instructions.

This study contributes to a deeper understanding of how to effectively leverage MLLMs by offering an equitable comparison of instruction languages. The main contributions of this study are as follows:

- We conduct a fair comparison by instructing MLLMs in English or target-language, eliminating the influence of translationese.
- Our primary findings indicate that instructions given in a particular language excel on respective tasks. Generally, target-language instructions outperform in lexical simplification tasks, while English instructions are more effective in reading comprehension tasks. Specifically, for classification tasks, instructions that align with the classification label's language tend to yield better performance.
- Our secondary findings highlight differences in MLLMs' features of generated texts and their instruction-following abilities under English versus target-language instructions. Notably, MLLMs adhere more closely to English instructions, regardless of effectiveness.

2 Related Work

Prompts for instruction-tuned models generally contain both instances and instructions. The study on whether MLLMs should be provided prompts in English or target-language can be categorized into instance-based and instruction-based approaches.

The instance-based approach focuses on translating instances into English. Huang et al. (2023) and Etxaniz et al. (2024) reported the effectiveness of having the LLM itself translate instances into English and then process them. Conversely, Intrator et al. (2024) reported translating instances into English led to a decrease in performance for PaLM2.

In contrast, the instruction-based approach, to which this study belongs, focuses on the language used for the instructions or prompt templates while keeping the instances unchanged. Lin et al. (2022), Muennighoff et al. (2023) and Ahuja et al. (2023) reported the effectiveness of English instructions and prompt templates, even for non-English tasks. However, these studies used multilingual datasets translated from English, such as XNLI (Conneau et al., 2018), as test data or target-language instructions translated from English, without considering the influence of translationese. On the other hand, Bareiß et al. (2024) used datasets not based on English but differed from our study by employing prompt templates based on translations and focusing on encoder-only models.

3 Methodology

In this section, we describe the methodology and experimental setup to conduct a fair comparison between English and target-language instructions in MLLMs, eliminating translationese effects.

3.1 Fair Instruction Construction

To ensure a fair comparison between English and target-language instructions, it is essential that both instructions convey the same content and are fluent enough. In our study, we create such instructions through a human-in-the-loop approach, which we refer to as "human-in-the-loop instruction construction." This approach involves the following steps, summarized in Figure 2:

Step 1. Manually defining the content to be included in the instructions for each task. These definitions serve as guidelines containing the key information necessary to perform the task and are not subject to translation in the following steps.

Step 2. Generating instructions in each language using GPT-4 based on the definitions created in Step 1. The instructions are generated independently in each language.

Step 3. Verifying whether the English and target-language instructions convey the same content with GPT-4. If differences are found, we repeat Step 2.

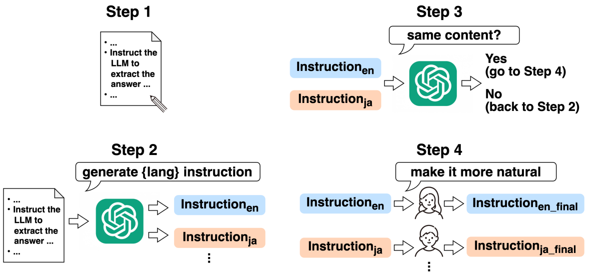


Figure 2: Overview of fair instruction construction.

Step 4. Having native speakers of each language refine the instructions to ensure natural phrasing and fluency.

We also considered having native speakers directly create instructions for each language based on the definitions from Step 1. However, this approach resulted in inconsistencies in content and style across languages. On the other hand, our construction process ensures that the instructions in each language convey the same content and are expressed in a linguistically natural manner.

The final instructions are listed in Appendix E.

3.2 Multilingual Testbenches

In this section, we describe the tasks conducted in this study and the test datasets, which were not derived from translation [3]. We provide examples of the instance for each task in Appendix B.2.

Lexical Simplification Task Lexical simplification (LS) is a task that involves simplifying a sentence by replacing a target word with a simpler synonym. For each instance, we generate a single, simpler synonym and measure accuracy based on whether the generated synonym is included in the gold-standard answer. The target-languages in the LS task are de, es, fr, ja, and zh. As test datasets, we use MultiLS (Shardlow et al., 2024) (de, es, fr, ja) and Chinese-LS (Qiang et al., 2023) (zh).

Machine Reading Comprehension Task Machine reading comprehension (MRC) is a task that involves answering a question based on a reference text. We extract an answer to the question from the reference text and measure accuracy based on whether the extracted answer exactly matches the gold-standard answer. The target-languages in the MRC task are de, es, fr, id, ja, ko, and zh. As test datasets, we use GermanQuAD (Möller et al., 2021) (de), SQAC (Gutiérrez-Fandiño et al., 2022) (es), FQuAD (d'Hoffschmidt et al., 2020) (fr), TyDiQA-Gold (Clark et al., 2020) (id, ja, ko), and DRCD (Shao et al., 2019) (zh).

Review Classification Task We perform a review classification (RC) task, which is a binary classification of whether a review sentence has a positive or negative rating. We consider two label settings—using English labels ('good-bad') and target-language labels [4]—and compare the macro-F1 between English and target-language instructions for each setting. The target-languages in the RC task are de, es, fr, id, ja, ko, and zh. As test datasets, we use MARC (Keung et al., 2020) (de, es, fr, ja, zh), NSMC (Park, 2015) (ko), and PRDECT-ID (Sutoyo et al., 2022) (id).

3.3 Multilingual LLMs

In this study, we primarily focus on instruction-tuned models. We conduct experiments using three open-source MLLMs: suzume (Devine, 2024) 8B, qwen2-instruct (Yang et al., 2024) 7B, and mistral-nemo-instruct (MistralAI, 2024) [5] 12B. These models are multilingual instruction-tuned versions of base models llama 3 (Dubey et al., 2024), qwen2, and mistral-nemo, respectively. Hereafter, we refer to these instruction-tuned models as llama3-i, qwen2-i, and mistraln-i. Appendix C.1 reports additional results for base models.

4 Results

Table 1 presents the experimental results across all target-languages for each task in zero-shot settings.

Lexical Simplification Task The experimental results indicate that target-language instructions tend to outperform English instructions in the LS task. Additionally, in Japanese, the performance of instructions translated from English significantly decreased because the numerical information contained in the English instructions was lost in translation (Appendix A.1). This result indicates that comparisons between English instructions and target-language instructions translated from English, as in previous studies, may not always be fair. In such biased conditions, English instructions are unjustly evaluated as more effective.

[3] Appendix B.1 provides more detailed descriptions of each dataset and our preprocessing methods where applicable.

[4] Appendix B.5 shows target-language labels.

[5] Unlike other languages, there is no description that id was included in its training data at MistralAI (2024); therefore, we do not perform experiments on id for mistral-nemo-instruct.

Task	Inst	Performance		
		llama3-i	qwen2-i	mistraln-i
LS	en	26.95	44.38	48.68
	tgt	**28.31**	**46.52**	**52.78**
	tgt-mt	23.33	40.64	46.12
MRC	en	**25.47**	**32.33**	**39.48**
	tgt	20.07	22.19	31.47
	tgt-mt	18.01	18.47	32.91
RC (en label)	en	**87.66**	**90.58**	**89.15**
	tgt	77.57	90.56	80.47
	tgt-mt	83.96	88.82	79.06
RC (tgt label)	en	66.72	86.49	65.34
	tgt	**70.14**	**89.46**	**65.47**
	tgt-mt	69.22	81.58	61.17

Table 1: Comparison of experimental results between en (English), tgt (target-language) and tgt-mt (target-language translated from English using Bing Translator) instructions for each task. The evaluation methods for performance in each task are described in Section 3.2. We list average scores across all target-languages. We highlight the best results for each model and task in bold.

Machine Reading Comprehension Task The experimental results indicate that English instructions tend to outperform target-language instructions in the MRC task. This trend contrasts with the LS task, indicating that whether English or target-language instructions perform better varies depending on the task.

Review Classification Task The experimental results indicate that in settings with English classification labels, English instructions tend to outperform target-language instructions. Conversely, in settings with target-language labels, target-language instructions tend to outperform English instructions. These findings suggest that in classification tasks, the optimal language depends on the classification labels, and using instructions that align with the labels' language can enhance the performance of MLLMs.

5 Analysis

5.1 Generation from Fair Instruction

The percentage of instances where the generated texts are the same between using English and target-language instructions is approximately 30% for llama3-i, 37% for qwen2-i, and 48% for mistraln-i in the MRC task. These results show that more than half of the generated texts differ when given two instructions that convey the same content but are written in different languages. In this section, we analyze the features of text generated by MLLMs using either English or target-language instructions.

Task	Inst	llama3-i	qwen2-i	mistraln-i
LS	en	9.94	8.23	7.08
	tgt	7.13	6.43	6.22
MRC	en	4.33	4.36	2.98
	tgt	2.16	1.47	1.76

Table 2: Percentage of instances where MLLMs generate texts in a language other than the target-language. We list the average percentage across all target-languages.

English instructions more often lead to generating unrelated languages. To identify the language of the texts generated by MLLMs, we use FastText (Joulin et al., 2016). Following previous studies (Wenzek et al., 2020; Kojima et al., 2024), we use only language identification results with an identification confidence score above 50%. Table 2 shows the percentage of instances where MLLMs generate texts in a language other than the target-language. These results indicate that using English instructions more often leads to the generation of text in a language other than the target-language. This observation is similar to the findings of Marchisio et al. (2024). Specifically, we found that when using English instructions, llama3-i tended to generate in English, while qwen2-i tended to generate in Chinese. We describe the detailed distribution of language identification in Appendix D.

Target-language instructions more often lead to generating uninformative answers like "There is no information." In the MRC task, although an answer is always present in the reference text, MLLMs occasionally generate awkward texts like "There is no information on the question in the reference." We manually counted the instances where MLLMs generated such responses in the Japanese and Spanish datasets. Table 3 shows the number of these instances. These results indicate that using target-language instructions causes MLLMs to generate such texts more often than when using English instructions. Notably, in some instances, MLLMs generate such texts with target-language instructions, whereas they provide the correct answer with English instructions. This observation suggests that using English instructions is more effective in leveraging the reading comprehension capabilities of MLLMs.

Lang	Inst	llama3-i	qwen2-i	mistraln-i
es	en	0	1	0
	tgt	8	18	2
ja	en	3	5	0
	tgt	28	15	3

Table 3: Number of instances where MLLMs generate texts like "There is no information on the question in the reference." in Spanish and Japanese for the MRC.

Task	Inst	llama3-i	qwen2-i	mistraln-i
LS	en	**19.95**	**2.31**	**0.35**
	tgt	23.54	2.97	0.91
MRC	en	**45.57**	**37.49**	**27.34**
	tgt	61.14	58.33	46.90

Table 4: Percentage of instances where the MLLM do not follow each instruction. We list the average percentage across all target-languages. The results for each language are in Table 17 in the Appendix.

5.2 Instruction-following Ability

We analyze the differences in the MLLMs' instruction-following ability between using English and target-language instructions by counting instances where MLLMs do not follow each instruction. We define a generated text as not following the instructions in the LS task if it contains more than five words [6] for de, es, fr, and zh and more than seven words [7] for ja as determined by spaCy (Honnibal et al., 2020). In the MRC task, we consider a generated text as not following the instructions if it contains any string not present in the reference text. Table 4 shows the percentage of instances where MLLMs do not follow each instruction. These results indicate that MLLMs follow English instructions more closely than target-language instructions. This observation suggests that using instructions in English is more effective for tasks requiring complex guidance.

5.3 Instruction Cross-Lingual Consistency

Qi et al. (2023) introduced cross-lingual consistency (CLC) and highlighted the importance of providing consistent user experiences when using the same LLM in different languages. However, as demonstrated in Sections 4 and 5.1, MLLMs often generate different outputs when given two instructions that convey the same content but are written in different languages. This difference indicates a low level of instruction CLC. To address this issue, we propose a few-shot approach that includes providing both task instructions and examples. We reveal that adopting a few-shot approach significantly enhances instruction CLC (Appendix C.3).

6 Conclusion

In this study, we conducted a fair comparison between English and target-language instructions for MLLMs, eliminating the influence of translationese. We revealed that whether English or target-language instructions tend to perform better depends on the task and classification labels. Additionally, we demonstrated that MLLMs exhibited differences in the features of generated texts and their instruction-following abilities when using English and target-language instructions.

Limitations

While we achieved a fair comparison between English and target-language instructions by employing datasets and instructions not based on translation from English, the range of languages and tasks we examined is limited. This is due to the fact that many multilingual datasets are created through translating from English, and a few datasets are independent of such translation. Furthermore, our study is currently restricted to high-resource languages, as non-translated datasets for low-resource languages are scarce, and finding native speakers to refine instructions in these languages is difficult. Investigating the features of tasks where English instructions perform better and those where target-language instructions perform better remains challenging, as it requires a wide variety of target-language datasets that are not based on translation.

Moreover, we used three state-of-the-art open-source MLLMs because the latest models have been shown to exhibit higher performance and superior instruction-following ability. However, many of the latest MLLM developers do not disclose key information, such as the distribution of languages in their training data. As a result, we were unable to conduct an analysis from the perspective of MLLMs' training data, such as analyzing why llama3-i tends to generate English while qwen2-i tends to generate Chinese.

Acknowledgements

This work was partly supported by JST, PRESTO Grant Number JPMJPR2366, Japan.

[6] We follow Lin et al. (2012) to filter generated texts that sound more like a sentence than a word or phrase.

[7] We follow Kudo and Kazawa (2009).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for NLI-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, volume 17 of *WWW ’ 24*, page 1318–1326. ACM.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. *Preprint*, arXiv:2405.12612.

Martin d’Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine

Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sauleh Eetemadi and Kristina Toutanova. 2014. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 159–164, Doha, Qatar. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, page 39–60.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 829–844, Mexico City, Mexico. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *Preprint*, arXiv:1612.03651.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Taku Kudo and Hideto Kazawa. 2009. Japanese web n-gram version 1. Linguistic Data Consortium.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. *Preprint*, arXiv:2406.20052.

MistralAI. 2024. Mistral NeMo. `https://mistral.ai/news/mistral-nemo/`.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Ayana Niwa and Hayate Iso. 2024. AmbigNLG: Addressing task ambiguity in instruction for NLG. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10733–10752, Miami, Florida, USA. Association for Computational Linguistics.

Lucy Park. 2015. Naver sentiment movie corpus v1.0. `https://github.com/e9t/nsmc`.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Jipeng Qiang, Kang Liu, Ying Li, Yun Li, Yi Zhu, Yun-Hao Yuan, Xiaocheng Hu, and Xiaoye Ouyang. 2023. Chinese lexical substitution: Dataset and method. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 29–42, Singapore. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. DRCD: a chinese machine reading comprehension dataset. *Preprint*, arXiv:1806.00920.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.

Rhio Sutoyo, Said Achmad, Andry Chowanda, Esther Widhi Andangsari, and Sani M. Isa. 2022. PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks. *Data in Brief*, 44:108554.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P. Xing, and Zhiting Hu. 2024. PromptAgent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*. ICLR 2024.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Shuly Wintner. 2016. Translationese: Between human and machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 18–19, Osaka, Japan. The COLING 2016 Organizing Committee.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

A Examples of Translationese

In this section, we present examples of translationese in both instructions and datasets in Japanese.

A.1 Instructions

We confirmed the negative impact of translationese in this study. A portion of English instructions and a portion of Japanese instructions translated from English in the LS task are listed under ID 1 in Table 5. The English instruction contains the quantifier ‘a,’ which indicates the generation of a single synonym. However, in the translated Japanese instructions, this quantifier was lost in translation. As a result, it became unclear whether MLLMs should generate a single synonym or multiple synonyms when using the translated Japanese instructions. Consequently, MLLMs often generated multiple synonyms, such as the five words ‘交番車, 車両, 付近の警備車, 駆けつけ車, 警察車’ for the target word ‘パトカー.’ This led to a significant decline in performance when using Japanese

ID	Original English sentence	Translated Japanese sentence
1	Please generate **a** simpler Japanese synonym for the word.	より簡単な日本語の同義語を生成してください。
2	You are an AI assistant whose purpose is to perform open-domain commonsense causal reasoning. You will be provided a premise and two **alternatives**, where the task is to select the **alternative** that more plausibly has a causal relation with the premise. ...	あなたは、オープンドメインの常識的な因果推論を実行することを目的としたAIアシスタントです。前提と2つの**選択肢**が提供され、その課題は、前提と因果関係を持つ**代替案**を選択することです。 ...
3	Sentence 1: It will be high with a long wall and capacity . Sentence 2: It will be high , with a long wall and a capacity .	Sentence 1: 長い壁と容量を伴う高いものとなるでしょう。 Sentence 2: それは高いところにあり、壁が長く、収容人数が多いでしょう。
4	Besides Kuykendall , Robert White and Joshua Soule Zimmerman served as Chancery Commissioner for Hampshire County .	カイケンデールに加えて、ロバート・ホワイトとジョシュア・スール・ジンマーマンがハンプシャー郡の衡平法裁判所コミッショナーを務めました。

Table 5: Examples of translationese in Japanese.

instructions translated from English, as shown in tgt-mt in Table 8.

Similarly, Ahuja et al. (2023) used Bing Translator to translate the English instructions into the target-language instructions. In their paper, they provided only the English instructions, not the non-English ones; therefore, we translated their English instructions into Japanese. The instructions used for Commonsense Reasoning tasks are listed under ID 2 in Table 5. In the English instruction, the term 'alternative' is used in the sense of 'option.' However, in the translated Japanese instruction, the first term of 'alternative' is expressed as '選択肢 (option)', while the second term is expressed as '代替案 (substitute).' This inconsistency causes the Japanese instruction to lack clarity and fluency, making it difficult to understand.

A.2 Datasets

PAWS-X (Yang et al., 2019) is a dataset for the Paraphrase Identification Task and is a multilingual dataset translated from English. Notably, the test data has been translated manually. Instances where two sentences are identified as paraphrases are listed under ID 3 in Table 5. In the Japanese instance, sentence 1 is either impossible to interpret or extremely difficult to understand. As a result, the two sentences of the Japanese instance cannot be considered a paraphrase.

Additionally, instances that differ from natural Japanese are listed under ID 4 in Table 5. The translated Japanese instance contains many transliteration [8], resulting in a style that differs from that of natural Japanese sentences.

These examples demonstrate that even in Japanese, a relatively high-resource language, the influence of translationese can be significant. Therefore, it is likely that languages with lower resources are even more affected by translationese. Based on this, we argue that the use of target-language datasets and instructions translated from English, as seen in previous studies, does not allow for a fair comparison between English and target-language instructions.

B Experiment Details

B.1 Test Datasets

We describe the datasets used in each task that are not based on translations from English.

Lexical Simplification Task For de, es, fr and ja, we use the MultiLS (Shardlow et al., 2024). MultiLS is a multilingual corpus of LS. This corpus has a test set of approximately 570 instances for each language. For zh, we use the Chinese-LS (Qiang et al., 2023). Chinese-LS is a Chinese corpus of LS. This corpus has 524 instances. We randomly sample 90% of the instances from the corpus as the test set, and use the remaining instances as the example set for few-shot settings.

Machine Reading Comprehension Task For de, we use the GermanQuAD (Möller et al., 2021). GermanQuAD is a German corpus and has a test set of 2,204 instances. For es, we use the SQAC (Gutiérrez-Fandiño et al., 2022). SQAC is

[8] Transliteration in Japanese is typically written in katakana.

Task	Test Instance	Answer
LS	Sentence: After the war, Hitler remained in the army and after receiving intelligence and oratory training, became an intelligence official tasked with infiltrating political parties and reporting to his superiors on their activities. Target word: infiltrating	invading, penetrating, intruding, entering, ...
MRC	Reference: Television formats portraying ordinary people in unscripted situations are almost as old as the television medium itself. Producer-host Allen Funt's Candid Camera, in which unsuspecting people were confronted with funny, unusual situations and filmed with hidden cameras, first aired in 1948, and is often seen as a prototype of reality television programming.[2][3] Question: What is considered the first reality TV show?	Candid Camera
RC	Two of the glasses were broken when I opened the package. Could you please be careful for packaging glass items.	bad

Table 6: Examples of instances from each task in English.

a Spanish corpus and has a test set of 1,910 instances. For fr, FQuAD (d'Hoffschmidt et al., 2020). FQuAD is a French corpus and has a valid set of 3,188 instances. For id, ja, and ko, we use the TyDiQA-Gold (Clark et al., 2020). TyDiQA-Gold is a multilingual corpus. This corpus has a valid set of 565 instances in id, 455 in ja, and 276 in ko. For zh, we use the DRCD (Shao et al., 2019). DRCD is a Chinese corpus and has a test set of 3,493 instances.

Review Classification Task For de, es, fr, ja, and zh, we use the MARC (Keung et al., 2020). MARC is a multilingual corpus of Amazon reviews of customers. This corpus has a test set of 5,000 reviews for each language, with ratings classified from 1 to 5. We use 4,000 reviews classified as positive or negative for each language. For ko, we use the NSMC (Park, 2015). NSMC is a Korean corpus of movie reviews from NAVER Movies. This corpus has 50,000 reviews classified as positive or negative. We randomly select 2,000 positive and 2,000 negative reviews as a test set. For id, we use the PRDECT-ID (Sutoyo et al., 2022). PRDECT-ID is an Indonesian corpus of product reviews from Tokopedia. This corpus has a test set of 5,400 reviews with ratings classified from 1 to 5. We perform downsampling to ensure an equal number of positive and negative reviews, and use 4,010 reviews classified as positive or negative.

B.2 Instance Examples from Each Task

In this study, we use target-language datasets and no English datasets. However, we provide examples of instances in English to ensure clarity for all readers of this paper. Table 6 lists instance examples from each tasks in English.

B.3 Text Generation

In this section, we describe the hyper-parameters and post-processing steps used during generation in both the LS and MRC tasks. The hyper-parameters of generation are as follows:

- temperature: 0.6
- top_p: 0.9
- max_new_tokens: 30

Following Iyer et al. (2023) and Wei et al. (2023), we extract the part of the text generated by MLLMs before the first EOS token or newline character as the output.

B.4 Label Selection

Many previous studies (Lin et al., 2022; Tanwar et al., 2023; Etxaniz et al., 2024) have used the label with the highest probability for the prompt in the classification label space as the LLM's prediction in classification tasks. Following these studies, in the RC task, we use the label with the highest probability as the next token after the input prompt for an MLLM's prediction.

B.5 Label Sets

Table 7 lists classification label sets for each language used in the target-language label setting of the RC task.

lang	good	bad
de	gut	schlecht
es	bueno	malo
fr	bon	mauvais
id	baik	buruk
ja	良い	悪い
ko	좋음	나쁨
zh	好	差

Table 7: Labels for each language used in the target-language label setting of the RC task.

B.6 Models

We conducted experiments with llama3-i [9], qwen2-i [10], mistraln-i [11], ayae-i [12], llama3-b [13], qwen2-b [14], and mistraln-b [15] from huggingface and used Quadro RTX 8000 in the all experiments. Llama3-i and llama3-b are published under the Llama 3 Community License Agreement. Qwen2-i, qwen2-b, mistraln-i and mistraln-b are published under the Apache License Version 2.0. Ayae-i are published under the Creative Commons Attribution-NonCommercial 4.0 International License.

C Additional results

C.1 Base models

We primarily focus on instruction-tuned models but also conduct experiments with base models. Hereafter, we refer to the base models llama3, qwen2, mistral-nemo as llama3-b, qwen2-b, and mistraln-b, respectively.

Tables 8, 9, 10 and 11 list the results for each language in the LS task, the MRC task, the RC task (English labels), and the RC task (target-language labels), respectively. In the LS task, target-language instructions tend to outperform English instructions for the base models, similarly to the instruction-tuned models. In the MRC task, English instructions tend to outperform target-language instructions for the base models other than llama3-b, similar to the instruction-tuned models. In the RC task with English classification labels, target-language instructions tend to outperform English instructions for the base models, unlike the instruction-tuned models. Notably, when using English instructions for the base models, the predictions are heavily skewed towards 'good'—phenomena referred to as *label bias* (Reif and Schwartz, 2024). For this issue, target-language instructions have the effect of mitigating the bias towards 'good' in the base models. In the RC task with target-language labels, whether English or target-language instructions perform better varies for each target-language.

C.2 Additional instruction-tuned model

Aya-Expanse is a MLLM released in October 2024, demonstrating superior multilingual performance compared to other MLLMs (Dang et al., 2024). Given its strong multilingual capabilities, we conduct experiments using Aya-Expanse 8B as an additional instruction-tuned model, which we refer to as ayae-i.

Tables 8, 9, 10 and 11 include the results of ayae-i for each task. These results indicate that ayae-i follows the same trend as other instruction-tuned models, where target-language instructions tend to achieve higher performance than English instructions in the LS task and the RC task (target-language labels), whereas English instructions tend to yield better performance in the MRC task and the RC task (English labels).

C.3 Few-shot Setting

We primarily focus on the zero-shot setting but also conduct experiments in the few-shot setting. For the few-shot examples, in the LS task, we randomly select four examples from the trial data for each test instance. For the MRC task, we select one example each for the questions 'who,' 'where,' 'what,' 'when,' and 'how' from the train data. In the RC task, we randomly select two reviews with a 'bad' label and two with a 'good' label from the training data for each test instance. Therefore, the LS and RC tasks are conducted in a 4-shot setting, while the MRC task is conducted in a 5-shot setting.

Tables 12, 13, 14 and 15 present the few-shot results for the LS task, the MRC task, the RC task (English labels), and the RC task (target-language labels), respectively. These results indicate that,

[9] https://huggingface.co/lightblue/suzume-llama-3-8B-multilingual

[10] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[11] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[12] https://huggingface.co/CohereForAI/aya-expanse-8b

[13] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[14] https://huggingface.co/Qwen/Qwen2-7B

[15] https://huggingface.co/mistralai/Mistral-Nemo-Base-2407

compared to the zero-shot setting, the performance differences between the different instructions are smaller in the few-shot setting.

Additionally, we investigate the percentage of instances where the generated text is identical when using English instructions and target-language instructions in the LS and MRC tasks, under the zero-shot and few-shot settings. Table 16 shows the percentage of instances where the texts generated by MLLMs are identical between using English and target-language instructions. This result indicates that in the zero-shot setting, the texts generated by MLLMs differ considerably between using English and target-language instructions, whereas in the few-shot setting, the number of identical texts increases significantly.

These findings reveal that in the zero-shot setting, even when English and target-language instructions convey the same content, MLLMs often generate different outputs, leading to a low instruction CLC. Adopting the few-shot approach can address this issue, significantly improving the consistency of generated texts across instructions in different languages, thereby greatly enhancing instruction CLC.

D Distributions of Languages in Generated Texts

Tables 18 and 19 show the language distribution of the generated texts identified by FastText when using English or target-language instructions in both the LS and MRC tasks. In the MRC task, we observed that using English instructions led to generating English text across all models. Additionally, for qwen2-i, even when the target-language was an alphabet-based language like `es`, using English instructions significantly increased the generation of Chinese text. For example, in the Spanish MRC task, qwen2-i generated ‘五个小时’ with English instructions, while the correct answer was ‘cinco horas.’

E Instructions

E.1 Construction Details

In steps 2 and 3 of the instruction construction process (Section 3.1), we used `gpt-4o-2024-05-13`. As native speakers in Step 4, we requested students pursuing a Doctors in NLP for `id` and `ko`, a student pursuing a Masters in NLP for `ja`, and an assistant professor in NLP for `zh`. For other languages, we recruited native speakers through the crowdsourcing platform Prolific [16].

E.2 Lexical Simplification Task

German

> Ich gebe Ihnen jetzt einen Satz und ein darin enthaltenes Wort.
> Bitte generiere ein einfacheres deutsches Synonym für das Wort.
> Generiere nur das Synonym und nichts anderes.
>
> Satz: `{sentence}`
>
> Wort: `{word}`
>
> Synonym:

English

> I will provide a sentence and a word included in the sentence.
> Please generate a simpler `{target language}` synonym for the word.
> Generate nothing but the synonym.
>
> Sentence: `{sentence}`
>
> Word: `{word}`
>
> Synonym:

Spanish

> Te proporcionaré una oración y una palabra de ella.
> Genere un sinónimo en español más sencillo para esta palabra.
> Genere solamente el sinónimo.
>
> Oración: `{sentence}`
>
> Palabra: `{word}`
>
> Sinónimo:

French

> Je vais vous donner une phrase et un mot tiré la phrase.
> Veuillez générer un synonyme en français plus simple pour le mot tiré.
> Ne générez que le synonyme.
>
> Phrase: `{sentence}`
>
> Mot: `{word}`
>
> Synonyme:

[16] https://www.prolific.com/

Japanese

これから文とその文に含まれる単語を与えます。
与えられた単語に対して、より簡単な日本語の同義語を一つ生成してください。
同義語以外は何も生成しないでください。

文: {sentence}

単語: {word}

同義語:

Chinese

我会给出一个句子并指定其中的一个词。
请生成一个该词的更简单的中文同义词。
只需生成同义词，不要生成其他内容。

句子: {sentence}

词: {word}

同义词:

E.3 Machine Reading Comprehension Task

German

Ich gebe Ihnen jetzt eine Frage und einen Referenzsatz.
Extrahiere die Antwort auf die Frage aus dem Referenzsatz.
Generiere nichts außer der Antwort.

Frage: {question}

Referenzsatz: {reference}

Antwort:

English

I will provide a question and a reference sentence.
Please extract the answer to the question from the reference sentence.
Generate nothing but the answer.

Question: {question}

Reference: {reference}

Answer:

Spanish

Te proporcionaré una pregunta y una oración de referencia.
Extraiga la respuesta a la pregunta de la oración de referencia.
Genere únicamente la respuesta.

Pregunta: {question}

Referencia: {reference}

Respuesta:

French

Je vais donner une question et une phrase de référence.
Veuillez extraire la réponse à la question à partir de la phrase de référence.
Ne générez rien d’autre que la réponse.

Question: {question}

Référence: {reference}

Réponse:

Indonesian

Saya akan memberikan sebuah pertanyaan dan sebuah kalimat referensi.
Silakan ekstrak jawaban untuk pertanyaan tersebut dari kalimat referensi.
Hasilkan hanya jawaban tanpa tambahan informasi lain.

Pertanyaan: {question}

Referensi: {reference}

Jawaban:

Japanese

これから質問と参照文を与えます。
質問に対する答えを参照文から抽出してください。
答え以外は生成しないでください。

質問: {question}

参照文: {reference}

答え:

Korean

지금부터 질문과 참고 문서를 입력합니다.
질문에 대한 답변을 참고 문서에서 추출해 주세요.
답변에 해당되는 부분만 생성해 주세요.

질문: {question}

참고 문서: {reference}

답변:

Chinese

我会提供一个问题和一段参考。
请根据这段参考，提取答案，回答问题。
请只生成答案。

问题: {question}

参考: {reference}

答案:

E.4 Review Classification Task

German

Ich gebe Ihnen eine Rezension.
Bitte bewerten Sie die Rezension anhand der folgenden Kriterien.
Wählen Sie ‘{label_good}’, wenn die Rezension eine positive Bewertung darstellt, und ‘{label_bad}’, wenn sie eine negative Bewertung darstellt.

Rezension: {sentence}

Bewertung:

English

I will provide a review.
Please rate the given review based on the following criteria.
Choose ‘{label_good}’ if the review indicates a high evaluation and ‘{label_bad}’ if it indicates a low evaluation.

Review: {sentence}

Rating:

Spanish

Voy a proporcionarte una reseña.
Por favor, califícala proporcionadamente según los siguientes criterios.
Elige ‘{label_good}’ si la reseña muestra una alta valoracion y ‘{label_bad}’ si es una baja valoración.

Reseña: {sentence}

Calificación:

French

Je vais fournir une critique.
Merci d’évaluer la critique en fonction des critères suivants.
Choisissez ‘{label_good}’ si la critique est positive et ‘{label_bad}’ si elle est négative.

Critique: {sentence}

Évaluation:

Indonesian

Saya akan memberikan sebuah ulasan.
Tolong nilai ulasan yang diberikan berdasarkan kriteria berikut.
Pilih ‘{label_good}’ jika ulasan menunjukkan evaluasi tinggi dan ‘{label_bad}’ jika menunjukkan evaluasi rendah.

Ulasan: {sentence}

Nilai:

Japanese

これからレビューの文を与えます。
そのレビューを以下の基準に基づいて評価してください。
そのレビューが高い評価を示す場合は‘{label_good}’を、低い評価を示す場合は‘{label_bad}’を選んでください。

レビュー: {sentence}

評価:

Korean

지금부터 리뷰를 입력합니다.
주어진 리뷰를 다음 기준에 따라 평가해 주세요.
리뷰가 높은 평가를 나타내는 경우 ‘{label_good}’을, 낮은 평가를 나타내는 경우 ‘{label_bad}’을 선택해 주세요.

리뷰: {sentence}

평가:

Chinese

我将提供一条评论。
请根据以下标准对给定的评论进行评分。
如果评论表示高度评价，请选择‘{label_good}’；如果评论表示不好的评价，请选择‘{label_bad}’。

评论: {sentence}

评分:

Target-lang	Instruct	Instruction-tuned model				Base model		
		llama3-i	qwen2-i	mistraln-i	ayae-i	llama3-b	qwen2-b	mistraln-b
de	en	**23.86**	37.19	29.12	50.35	28.07	15.61	1.23
	tgt	22.46	**37.37**	**35.61**	**53.33**	**29.30**	**22.46**	**6.49**
	tgt-mt	19.65	38.60	18.77	50.70	30.00	24.04	0.18
es	en	44.01	60.54	58.35	**64.42**	**49.58**	34.23	**6.24**
	tgt	**48.06**	**62.56**	**66.44**	57.67	26.14	**53.63**	1.85
	tgt-mt	34.91	63.91	68.80	64.25	32.72	54.13	0.84
fr	en	28.82	57.47	58.70	53.25	39.54	24.96	7.21
	tgt	**30.40**	**65.91**	**62.74**	**61.15**	**43.94**	**54.13**	**19.51**
	tgt-mt	31.81	55.89	56.24	60.10	32.86	44.46	11.60
ja	en	15.96	**26.67**	37.72	44.91	12.98	18.07	32.46
	tgt	**17.02**	26.49	**38.07**	**45.43**	**13.86**	**19.30**	**35.44**
	tgt-mt	7.54	4.74	24.21	26.84	5.96	4.04	11.75
zh	en	22.10	40.04	59.52	57.77	16.41	32.39	45.51
	tgt	**23.63**	**40.26**	**61.05**	**59.96**	**17.72**	**32.82**	**56.24**
	tgt-mt	22.76	40.04	62.58	33.92	13.79	27.13	52.52

Table 8: Experimental results of the zero-shot setting in the LS task. We highlight the higher score between ‘en’ and ‘tgt’ in bold.

Target-lang	Instruct	Instruction-tuned model				Base model		
		llama3-i	qwen2-i	mistraln-i	ayae-i	llama3-b	qwen2-b	mistraln-b
de	en	**12.75**	**24.95**	**26.86**	**35.48**	10.21	**12.79**	13.16
	tgt	9.80	16.61	26.04	30.76	**16.38**	12.02	**13.88**
	tgt-mt	8.67	13.70	24.41	25.36	11.93	11.43	12.70
es	en	**20.37**	**25.13**	**25.65**	**34.14**	13.09	**13.61**	**12.25**
	tgt	15.81	13.66	17.75	25.13	**13.14**	9.16	9.90
	tgt-mt	17.80	16.54	17.64	25.34	18.12	10.79	8.74
fr	en	12.95	**23.90**	**29.23**	**34.69**	10.19	**14.37**	**14.59**
	tgt	**16.66**	14.77	23.12	26.07	**13.80**	12.77	13.61
	tgt-mt	16.59	12.30	25.22	23.84	11.89	13.14	11.20
id	en	**34.69**	**46.90**	–	**57.70**	22.48	**37.52**	–
	tgt	33.98	42.48	–	53.63	**30.27**	33.27	–
	tgt-mt	23.54	31.86	–	33.98	24.78	30.27	–
ja	en	**43.52**	**41.76**	**58.02**	**64.84**	28.57	**41.10**	**39.34**
	tgt	33.19	38.02	52.53	63.52	**39.78**	30.77	35.60
	tgt-mt	27.91	23.30	45.71	55.82	32.75	29.45	27.47
ko	en	**25.72**	**40.94**	**55.43**	**59.42**	**21.01**	**36.96**	**30.80**
	tgt	2.54	7.97	34.42	12.68	18.84	13.41	13.77
	tgt-mt	1.81	10.14	43.84	46.74	22.46	16.30	24.64
zh	en	28.26	**22.70**	**41.68**	**57.91**	**14.86**	19.55	**30.15**
	tgt	**28.49**	21.84	34.98	46.78	11.65	**22.16**	26.40
	tgt-mt	29.77	21.41	40.65	48.04	17.64	19.98	28.54

Table 9: Experimental results of the zero-shot setting in the MRC task. We highlight the higher score between ‘en’ and ‘tgt’ in bold.

Target-lang	Instruct	Instruction-tuned model				Base model		
		llama3-i	qwen2-i	mistraln-i	ayae-i	llama3-b	qwen2-b	mistraln-b
de	en	**90.07**	92.31	**92.52**	**95.15**	42.96	36.37	80.87
	tgt	85.17	**93.92**	86.66	93.86	**52.63**	**85.99**	**93.82**
	tgt-mt	83.29	88.90	90.41	94.47	43.76	59.52	93.80
es	en	**90.45**	92.18	**91.47**	**94.62**	42.91	36.58	62.71
	tgt	74.23	**93.50**	79.41	94.57	**72.37**	**73.75**	**83.22**
	tgt-mt	84.97	91.11	52.67	93.87	72.06	39.17	70.01
fr	en	**89.73**	**92.84**	**91.72**	**94.80**	38.25	**37.10**	64.54
	tgt	72.62	92.74	78.25	91.12	**49.90**	33.50	**86.93**
	tgt-mt	87.97	88.06	88.10	92.48	83.86	50.98	82.54
id	en	**90.61**	**96.93**	–	97.81	36.89	36.89	–
	tgt	86.56	95.81	–	**98.13**	**73.86**	**43.93**	–
	tgt-mt	85.56	94.93	–	98.23	81.57	49.14	–
ja	en	**88.47**	**89.55**	**90.77**	**93.47**	39.36	47.82	71.90
	tgt	87.38	88.17	86.58	91.64	**82.56**	**69.06**	**86.64**
	tgt-mt	91.27	89.39	89.88	92.40	89.88	61.86	91.32
ko	en	**76.78**	**81.32**	83.99	**88.37**	44.90	33.78	42.54
	tgt	53.09	80.85	**85.30**	86.86	**50.26**	**33.89**	**63.45**
	tgt-mt	70.04	80.60	75.17	85.76	71.34	39.42	36.74
zh	en	**87.52**	88.92	**79.34**	**87.03**	34.05	64.18	80.71
	tgt	83.94	**88.96**	75.72	85.12	**39.37**	**89.90**	**82.54**
	tgt-mt	84.57	88.78	74.41	85.81	39.22	89.72	83.66

Table 10: Experimental results of the zero-shot setting with English labels in the RC task. We highlight the higher score between ‘en’ and ‘tgt’ in bold.

Target-lang	Instruct	Instruction-tuned model				Base model		
		llama3-i	qwen2-i	mistraln-i	ayae-i	llama3-b	qwen2-b	mistraln-b
de	en	33.44	61.42	58.75	89.03	33.33	33.33	33.33
	tgt	**50.60**	**78.05**	**59.91**	**94.60**	**33.44**	33.33	**34.33**
	tgt-mt	33.56	34.05	36.58	94.19	33.33	33.33	33.44
es	en	87.89	90.12	89.56	88.71	**82.74**	85.80	**90.66**
	tgt	**91.59**	**93.32**	**93.12**	**89.88**	76.12	**93.40**	85.79
	tgt-mt	87.11	87.62	92.59	87.36	79.79	58.67	77.50
fr	en	**39.86**	**92.82**	33.33	34.33	33.33	**87.52**	33.33
	tgt	33.67	92.19	33.33	**37.47**	33.33	82.55	33.33
	tgt-mt	38.26	90.57	33.33	36.05	33.33	52.80	33.33
id	en	81.03	**97.46**	–	96.03	**90.51**	70.05	–
	tgt	**93.59**	96.38	–	**97.21**	90.13	**92.38**	–
	tgt-mt	92.69	97.43	–	96.61	96.96	94.67	–
ja	en	82.20	91.31	**86.73**	**93.75**	33.33	**93.45**	**59.89**
	tgt	**83.06**	**91.91**	82.56	92.51	**36.15**	93.40	37.79
	tgt-mt	90.21	88.20	85.60	93.29	47.21	87.63	67.48
ko	en	**63.08**	82.99	86.67	**78.54**	33.33	**50.15**	34.11
	tgt	58.85	**84.91**	**87.02**	78.31	**33.78**	34.87	**40.92**
	tgt-mt	66.48	84.30	84.98	84.05	34.82	40.71	37.89
zh	en	79.55	89.30	**37.00**	82.25	84.90	**87.76**	33.33
	tgt	**79.64**	**89.47**	36.90	**86.44**	**85.68**	86.89	33.33
	tgt-mt	76.25	88.87	33.94	86.49	83.69	85.85	33.33

Table 11: Experimental results of the zero-shot setting with target-language labels in the RC task. We highlight the higher score between ‘en’ and ‘tgt’ in bold.

Target-lang	Instruct	Instruction-tuned model			Base model		
		llama3-i	qwen2-i	mistraln-i	llama3-b	qwen2-b	mistraln-b
de	en	27.19	32.63	50.53	9.47	26.67	41.05
	tgt	29.47	31.40	49.82	11.93	26.32	49.30
	tgt-mt	29.47	33.86	44.74	10.35	26.14	29.30
es	en	67.28	70.66	72.85	11.80	68.47	15.85
	tgt	63.58	71.16	75.21	6.58	71.16	16.02
	tgt-mt	64.92	73.19	73.52	8.26	75.89	7.42
fr	en	44.82	61.86	75.04	5.80	58.52	35.15
	tgt	46.75	63.44	72.41	8.08	63.27	55.36
	tgt-mt	44.99	64.15	72.23	4.57	61.16	61.51
ja	en	20.53	27.19	45.79	14.39	24.39	33.68
	tgt	21.75	24.91	43.68	14.39	21.58	39.82
	tgt-mt	17.19	26.49	42.28	13.68	21.75	36.84
zh	en	30.63	36.11	67.83	15.75	32.82	53.83
	tgt	32.17	36.11	66.74	17.07	33.26	58.64
	tgt-mt	29.76	34.14	68.27	16.19	35.45	57.77

Table 12: Experimental results of the few-shot setting in the LS task.

Target-lang	Instruct	Instruction-tuned model			Base model		
		llama3-i	qwen2-i	mistraln-i	llama3-b	qwen2-b	mistraln-b
de	en	15.65	30.67	27.59	8.44	18.65	0.86
	tgt	14.88	28.09	26.68	8.03	15.88	17.33
	tgt-mt	14.25	27.77	26.09	6.94	14.75	16.97
es	en	18.95	39.58	28.38	12.93	25.92	3.14
	tgt	15.76	36.39	18.90	15.39	23.09	3.14
	tgt-mt	15.92	36.75	19.95	14.61	22.46	3.61
fr	en	16.12	35.48	32.06	9.03	25.47	25.75
	tgt	15.81	35.45	31.68	10.16	24.65	27.92
	tgt-mt	15.12	36.04	31.37	10.19	23.90	27.29
id	en	29.03	59.29	–	21.95	39.47	–
	tgt	29.20	58.94	–	22.12	39.65	–
	tgt-mt	27.96	56.28	–	21.24	36.81	–
ja	en	50.77	53.19	60.44	39.12	46.81	45.71
	tgt	45.49	60.00	60.66	46.15	46.59	46.81
	tgt-mt	43.52	54.29	57.36	40.88	44.62	43.52
ko	en	30.80	56.88	50.36	26.45	43.12	35.14
	tgt	28.26	57.97	44.57	24.64	32.25	38.77
	tgt-mt	27.54	56.16	44.20	25.72	37.32	38.04
zh	en	30.23	31.89	47.64	16.40	23.05	32.52
	tgt	30.80	30.60	47.38	16.32	23.25	34.33
	tgt-mt	30.03	31.41	47.64	14.26	23.25	34.67

Table 13: Experimental results of the few-shot setting in the MRC task.

Target-lang	Instruct	Instruction-tuned model			Base model		
		llama3-i	qwen2-i	mistraln-i	llama3-b	qwen2-b	mistraln-b
de	en	90.37	92.51	94.00	86.22	35.57	78.96
	tgt	91.62	93.50	92.51	89.03	49.20	87.81
	tgt-mt	92.56	92.57	92.64	90.83	43.42	87.63
es	en	88.57	92.00	90.61	83.91	39.17	75.56
	tgt	87.24	92.03	87.64	86.02	39.97	78.40
	tgt-mt	88.65	91.95	77.24	88.69	35.36	69.57
fr	en	89.34	92.71	91.17	83.13	35.95	76.59
	tgt	90.25	93.35	89.89	87.38	34.49	88.14
	tgt-mt	89.09	91.88	89.64	84.69	36.83	85.36
id	en	94.48	98.03	–	91.88	34.21	–
	tgt	93.97	94.85	–	95.58	33.39	–
	tgt-mt	95.33	93.47	–	95.98	34.27	–
ja	en	84.93	88.74	91.49	81.99	40.21	92.37
	tgt	90.94	87.67	89.92	90.08	34.60	90.92
	tgt-mt	91.40	90.03	88.70	90.56	33.89	90.24
ko	en	75.05	81.16	88.07	66.37	33.56	74.25
	tgt	77.12	84.41	87.92	75.96	36.11	84.82
	tgt-mt	75.99	84.60	87.79	74.74	38.51	83.76
zh	en	83.56	88.92	84.65	76.66	49.83	81.34
	tgt	85.46	89.47	80.39	84.92	47.74	79.40
	tgt-mt	85.55	89.35	78.30	85.23	47.83	77.67

Table 14: Experimental results of the few-shot setting with English labels in the RC task.

Target-lang	Instruct	Instruction-tuned model			Base model		
		llama3-i	qwen2-i	mistraln-i	llama3-b	qwen2-b	mistraln-b
de	en	38.20	44.23	53.16	33.33	33.33	33.33
	tgt	33.33	43.89	48.01	33.33	33.33	33.83
	tgt-mt	33.33	42.89	39.22	33.33	33.33	33.44
es	en	91.18	91.04	93.74	78.29	87.96	91.50
	tgt	93.04	91.09	92.93	86.28	92.85	94.25
	tgt-mt	91.01	87.88	90.23	82.29	91.21	93.80
fr	en	77.17	90.37	33.33	33.33	90.92	33.33
	tgt	81.67	91.53	33.33	33.67	90.10	33.33
	tgt-mt	75.92	90.27	33.33	33.33	90.05	33.33
id	en	92.35	98.35	–	97.08	98.15	–
	tgt	89.77	97.86	–	96.61	98.23	–
	tgt-mt	86.18	97.58	–	97.93	98.10	–
ja	en	90.38	92.47	91.95	78.90	88.79	86.50
	tgt	90.82	90.39	91.49	80.19	85.20	88.50
	tgt-mt	91.67	88.95	89.03	84.41	83.19	89.39
ko	en	68.90	78.95	87.56	33.56	34.16	74.80
	tgt	69.34	82.47	86.85	38.95	33.83	84.23
	tgt-mt	70.52	79.89	87.47	41.79	33.99	80.24
zh	en	84.59	89.18	50.99	85.92	88.37	33.33
	tgt	78.59	89.50	44.03	84.19	88.05	33.33
	tgt-mt	74.71	89.82	43.43	80.17	87.67	33.33

Table 15: Experimental results of the few-shot setting with target-language labels in the RC task.

Target-lang	Shot	LS			MRC		
		llama3-i	qwen2-i	mistraln-i	llama3-i	qwen2-i	mistraln-i
de	zero	25.79	43.86	27.72	23.59	30.72	48.55
	few	50.88	52.46	55.09	42.24	58.67	51.54
es	zero	26.98	46.37	39.97	39.90	31.31	44.71
	few	48.90	57.00	60.54	43.61	65.39	46.86
fr	zero	17.22	46.92	40.07	27.23	31.65	47.02
	few	48.33	53.78	60.28	39.21	69.23	57.87
id	zero	–	–	–	40.18	57.52	–
	few	–	–	–	49.20	72.04	–
ja	zero	23.86	37.37	49.82	34.73	50.33	63.30
	few	36.49	42.81	49.65	55.16	73.85	71.65
ko	zero	–	–	–	4.35	16.30	37.32
	few	–	–	–	36.96	74.28	59.06
zh	zero	29.98	54.49	60.61	37.79	39.59	47.12
	few	47.26	62.36	71.55	46.15	55.00	73.95

Table 16: Percentage of instances where the texts generated by MLLMs are the same between using English and target-language instructions.

Target-lang	Inst	LS			MRC		
		llama3-i	qwen2-i	mistraln-i	llama3-i	qwen2-i	mistraln-i
de	en	24.04	1.40	0.70	48.23	32.30	28.13
	tgt	28.25	2.63	1.40	70.46	56.85	37.98
es	en	23.95	3.88	0.34	40.63	35.50	31.83
	tgt	13.83	2.87	0.34	47.38	66.91	50.89
fr	en	37.61	1.76	0.00	61.04	38.99	33.78
	tgt	38.49	1.05	0.18	59.41	67.69	49.91
id	en	–	–	–	41.24	30.44	–
	tgt	–	–	–	48.50	38.23	–
ja	en	5.79	0.88	0.00	36.70	37.80	12.38
	tgt	13.86	2.81	1.23	58.68	38.24	23.08
ko	en	–	–	–	53.62	38.04	19.20
	tgt	–	–	–	94.20	85.14	51.09
zh	en	5.03	1.53	0.00	37.53	49.36	20.18
	tgt	20.13	3.72	0.00	49.36	55.22	44.03

Table 17: Percentage of instances where the MLLM do not follow each instruction in each target-language.

Model	target-lang	Inst	Language identified by FastText									
			en	de	es	fr	id	ja	ko	zh	other	low
llama3-i	de	en	3.27	86.89	0.05	0.27	0.00	0.00	0.00	0.00	0.36	9.17
		tgt	3.22	85.84	0.05	0.23	0.00	0.00	0.00	0.00	0.27	10.39
	es	en	1.26	0.10	86.44	0.10	0.05	0.05	0.00	0.00	1.36	10.63
		tgt	0.37	0.05	88.53	0.16	0.00	0.00	0.00	0.00	0.89	10.00
	fr	en	6.49	0.09	0.00	66.19	0.00	0.00	0.00	0.00	1.04	26.19
		tgt	1.82	0.09	0.03	83.75	0.00	0.00	0.00	0.00	1.29	13.02
	id	en	3.36	0.53	0.00	0.00	74.34	0.00	0.00	0.18	2.30	19.29
		tgt	0.88	0.35	0.00	0.18	76.99	0.00	0.00	0.00	1.77	19.82
	ja	en	1.76	0.00	0.00	0.00	0.00	95.16	0.00	0.66	0.22	2.20
		tgt	0.00	0.00	0.00	0.00	0.00	99.78	0.00	0.00	0.00	0.22
	ko	en	1.45	0.00	0.36	0.00	0.00	0.00	92.75	0.36	0.36	4.71
		tgt	0.00	0.00	0.00	0.00	0.00	0.00	99.28	0.00	0.00	0.72
	zh	en	0.83	0.00	0.00	0.06	0.00	3.26	0.06	94.42	0.06	1.32
		tgt	0.26	0.03	0.00	0.03	0.03	3.01	0.03	95.68	0.14	0.80
qwen2-i	de	en	2.18	90.15	0.05	0.18	0.00	0.09	0.00	1.32	0.23	5.81
		tgt	0.95	95.42	0.00	0.27	0.00	0.00	0.00	0.09	0.05	3.22
	es	en	1.10	0.00	89.74	0.21	0.00	0.10	0.00	1.41	0.84	6.60
		tgt	0.68	0.00	93.98	0.16	0.00	0.00	0.00	0.05	0.68	4.45
	fr	en	1.82	0.19	0.06	90.65	0.00	0.28	0.00	1.38	0.47	5.14
		tgt	0.91	0.13	0.03	94.82	0.00	0.00	0.00	0.06	0.31	3.73
	id	en	1.24	0.00	0.00	0.00	89.38	0.18	0.00	0.53	1.59	7.08
		tgt	0.00	0.00	0.00	0.00	92.04	0.00	0.00	0.00	1.42	6.55
	ja	en	0.00	0.00	0.00	0.00	0.00	96.48	0.00	3.52	0.00	0.00
		tgt	0.00	0.00	0.00	0.00	0.00	99.78	0.00	0.22	0.00	0.00
	ko	en	0.72	0.00	0.00	0.00	0.00	1.45	92.03	5.43	0.00	0.36
		tgt	0.00	0.00	0.00	0.00	0.00	0.00	99.64	0.36	0.00	0.00
	zh	en	0.09	0.00	0.03	0.03	0.00	3.69	0.00	95.48	0.11	0.57
		tgt	0.06	0.00	0.00	0.00	0.00	3.49	0.03	95.79	0.34	0.29
mistraln-i	de	en	2.27	91.42	0.00	0.14	0.00	0.00	0.00	0.05	0.27	5.85
		tgt	1.54	92.15	0.00	0.09	0.00	0.00	0.00	0.05	0.36	5.81
	es	en	1.26	0.10	91.05	0.21	0.00	0.00	0.00	0.00	0.79	6.60
		tgt	0.63	0.00	90.58	0.42	0.00	0.00	0.00	0.00	0.73	7.64
	fr	en	1.51	0.13	0.09	91.50	0.00	0.00	0.00	0.00	0.85	5.93
		tgt	1.04	0.09	0.03	92.10	0.00	0.00	0.00	0.00	0.60	6.15
	ja	en	0.00	0.00	0.00	0.00	0.00	98.90	0.00	0.88	0.22	0.00
		tgt	0.00	0.00	0.00	0.00	0.00	99.34	0.00	0.22	0.00	0.44
	ko	en	0.72	0.00	0.00	0.00	0.00	2.54	94.57	1.45	0.36	0.36
		tgt	0.36	0.00	0.00	0.00	0.00	0.72	97.83	0.72	0.00	0.36
	zh	en	0.31	0.06	0.00	0.03	0.00	1.63	0.00	97.25	0.06	0.66
		tgt	0.09	0.03	0.00	0.00	0.00	1.75	0.00	97.17	0.20	0.77

Table 18: Distributions of languages in generated texts when using each instruction in the MRC task. ‘low’ indicates instances of generated text where the confidence score of language identification by FastText is less than 50%.

Model	target-lang	Inst	Language identified by FastText									
			en	de	es	fr	id	ja	ko	zh	other	low
llama3-i	de	en	0.88	74.21	0.00	0.00	0.00	0.00	0.00	0.00	4.56	20.35
		tgt	0.53	76.67	0.18	0.00	0.00	0.00	0.00	0.00	4.56	18.07
	es	en	3.04	0.17	67.28	0.34	0.00	0.00	0.00	0.00	3.54	25.63
		tgt	0.84	0.34	73.02	0.67	0.00	0.00	0.00	0.00	2.87	22.26
	fr	en	7.91	0.53	0.88	65.38	0.00	0.00	0.00	0.00	2.46	22.85
		tgt	3.87	0.00	1.05	58.00	0.00	0.18	0.00	0.00	2.64	34.27
	ja	en	0.53	0.18	0.00	0.35	0.00	92.28	0.35	5.09	0.53	0.70
		tgt	0.70	0.18	0.00	0.18	0.00	90.88	0.00	4.74	0.35	2.98
	zh	en	1.75	0.22	0.00	1.31	0.00	13.35	0.22	72.43	1.53	9.19
		tgt	1.75	0.00	0.00	0.88	0.00	7.00	0.00	73.74	2.19	14.44
qwen2-i	de	en	1.40	80.70	0.18	0.53	0.00	0.35	0.00	1.05	3.16	12.63
		tgt	0.88	83.68	0.00	0.00	0.00	0.00	0.00	1.05	2.28	12.11
	es	en	4.72	0.17	76.73	0.34	0.00	0.51	0.00	0.67	2.02	14.84
		tgt	0.84	0.17	81.28	0.00	0.00	0.00	0.00	0.67	1.85	15.18
	fr	en	6.15	0.53	0.18	79.44	0.00	0.35	0.00	1.05	1.41	10.90
		tgt	3.34	0.00	0.53	86.99	0.00	0.18	0.00	1.05	1.23	6.68
	ja	en	0.35	0.00	0.00	0.00	0.00	94.04	0.00	4.04	0.18	1.40
		tgt	0.00	0.00	0.00	0.00	0.00	92.46	0.00	5.96	0.53	1.05
	zh	en	1.09	0.66	0.00	1.09	0.00	7.66	0.66	81.62	0.66	6.56
		tgt	0.66	0.22	0.22	0.22	0.00	9.63	0.44	80.96	0.22	7.44
mistraln-i	de	en	2.98	70.53	0.35	0.53	0.00	0.53	0.00	0.18	2.28	22.63
		tgt	0.70	78.77	0.18	0.35	0.00	0.00	0.00	0.00	2.46	17.54
	es	en	2.02	0.00	72.18	0.84	0.00	0.00	0.00	0.34	2.19	22.43
		tgt	1.01	0.00	81.79	0.34	0.00	0.00	0.00	0.00	1.69	15.18
	fr	en	3.69	0.00	0.53	80.49	0.00	0.00	0.00	0.18	1.23	13.88
		tgt	4.57	0.35	0.00	82.95	0.00	0.00	0.00	0.00	1.05	11.07
	ja	en	0.35	0.00	0.00	0.18	0.00	91.40	0.18	6.14	0.18	1.58
		tgt	0.18	0.00	0.00	0.00	0.00	89.82	0.00	7.37	0.35	2.28
	zh	en	0.00	0.22	0.66	0.22	0.00	7.00	0.66	81.62	1.75	7.88
		tgt	0.00	0.22	0.66	0.44	0.00	5.69	0.66	84.25	2.84	5.25

Table 19: Distributions of languages in generated texts when using each instruction in the LS task. ‘low’ indicates instances of generated text where the confidence score of language identification by FastText is less than 50%.