Deep Exploration of Cross-Lingual Zero-Shot Generalization in Instruction Tuning

Janghoon Han[*] **Changho Lee**[*] **Joongbo Shin**
Stanley Jungkyu Choi **Honglak Lee** **Kyunghoon Bae**
LG AI Research
{janghoon.han,changho.lee}@lgresearch.ai

Abstract

Instruction tuning has emerged as a powerful technique, significantly boosting zero-shot performance on unseen tasks. While recent work has explored cross-lingual generalization by applying instruction tuning to multilingual models, previous studies have primarily focused on English, with a limited exploration of non-English tasks. For an in-depth exploration of cross-lingual generalization in instruction tuning, we perform instruction tuning individually for two distinct language meta-datasets. Subsequently, we assess the performance on unseen tasks in a language different from the one used for training. To facilitate this investigation, we introduce a novel non-English meta-dataset named "KORANI" (Korean Natural Instruction), comprising 51 Korean benchmarks. Moreover, we design cross-lingual templates to mitigate discrepancies in language and instruction-format of the template between training and inference within the cross-lingual setting. Our experiments reveal consistent improvements through cross-lingual generalization in both English and Korean, outperforming baseline by average scores of 20.7% and 13.6%, respectively. Remarkably, these enhancements are comparable to those achieved by monolingual instruction tuning and even surpass them in some tasks. The result underscores the significance of relevant data acquisition across languages over linguistic congruence with unseen tasks during instruction tuning[1].

1 Introduction

Recent studies have highlighted Instruction Tuning, where large language models are fine-tuned using instructions (templates) for various tasks, resulting in significant improvements in zero-shot performance on unseen tasks (Wei et al., 2022; Sanh et al., 2022; Wang et al., 2022; Chung et al., 2022; Luccioni et al., 2022; Ouyang et al., 2022; Zhong et al., 2021). Several works have investigated the effectiveness of instruction tuning to show cross-lingual zero-shot generalization. For example, Wang et al. (2022); Muennighoff et al. (2022); Jang et al. (2023); Li et al. (2023) apply instruction tuning to multilingual large language models (Conneau et al., 2020; Xue et al., 2021; Scao et al., 2022; Lin et al., 2022), pre-trained with multiple languages, and demonstrate cross-lingual generalization ability by achieving meaningful performance enhancements for unseen tasks in other languages.

However, prior investigations into cross-lingual generalization through instruction tuning, as observed in works by Wang et al. (2022) and Muennighoff et al. (2022), have predominantly focused on English and have limited diversity of tasks in non-English. This limited scope makes it challenging to thoroughly evaluate the effectiveness of cross-lingual zero-shot generalization, as direct comparisons with monolingual instruction tuning within the same language are not attainable. Moreover, previous research lacks validation of cross-lingual generalization of instruction tuning for languages other than English, and the evaluation of non-English datasets has been confined to specific tasks, offering only partial understanding.

To fill these gaps, our study undertakes a comprehensive investigation of the cross-lingual zero-shot generalization in instruction tuning. We define a new cross-lingual setting as the case where the training and inference language differs. In the setting, we instruction tune for two languages meta-dataset separately and evaluate the other language's unseen tasks. Specifically, to examine how effective cross-lingual zero-shot generalization is for a different language, we conduct a comparative analysis between cross-lingual instruction tuning and monolingual instruction tuning, where the latter indicates models trained and tested on the same

[*] indicates equal contribution.

[1] https://github.com/CHLee0801/KORANI-Instruction-Tuning

Findings of the Association for Computational Linguistics: ACL 2024, pages 15436–15452
August 11-16, 2024 ©2024 Association for Computational Linguistics

language's meta-dataset.

The collection of a non-English meta-dataset (Triantafillou et al., 2020) is imperative for the comprehensive examination of cross-lingual generalization of instruction tuning. However, collecting diverse supervised task datasets for non-English poses a substantial challenge due to the limited availability of open-source data in non-English languages compared to English. To address this issue, we propose a novel non-English language meta-dataset named **KORANI**, short for **KOR**e**A**n **N**atural **I**nstruction. This meta-dataset comprises 51 diverse Korean benchmarks, including 34 NLU benchmarks and 17 NLG benchmarks. Notably, KORANI surpasses the quantity of non-English benchmarks explored in previous multilingual research (Wang et al., 2022; Muennighoff et al., 2022) and approaches the size of P3 datasets (Sanh et al., 2022), which we employ as English benchmarks in our study.

Furthermore, in the cross-lingual setting, the language and instructional format of templates are different during the training and inference phase. These discrepancies in the template may contribute to suboptimal model performance (Muennighoff et al., 2022; Liang et al., 2023; Sun et al., 2024). To address this issue, we construct cross-lingual templates to align the template between the training and inference phases.

Our experiments show that cross-lingual instruction tuning consistently improves the zero-shot performance of unseen tasks in both English and Korean. Surprisingly, these cross-lingual performances are comparable to those of monolingual instruction-tuned models across various tasks, and they even surpass some tasks when cross-lingual templates are applied. These findings suggest that learning relevant tasks, even in different languages, is more crucial for performance improvement than ensuring linguistic congruence within unseen tasks in instruction tuning. Furthermore, our findings reframe the traditional view that cross-lingual instruction tuning merely enhances performance in low-resource languages, suggesting that it could serve as a viable alternative to monolingual instruction tuning.

Our contributions are summarized as follows:

- Our study enhances the understanding of cross-lingual instruction tuning by demonstrating that it can match the performance of monolingual tuning, emphasizing the importance of learning relevant tasks across languages.

- We introduce a new dataset called **KORANI**, comprising diverse Korean benchmarks, which provides a valuable resource for instruction tuning in non-English languages.

- We introduce cross-lingual templates to both the P3 dataset and KORANI, and confirm the robustness of cross-lingual zero-shot generalization achieved with these templates.

2 Related Work

2.1 Instruction Tuning

Instruction Tuning represents a learning methodology that enhances the zero-shot performance of unseen tasks by leveraging various Natural Language Processing (NLP) tasks. Instruction tuning explicitly trains NLP tasks using a multi-task training approach, and leverages templates to learn the salient characteristics of these tasks. By combining datasets and templates, instruction tuning induces robust generalization for unseen tasks with new templates adapted to assist the model for problem solving capability. Previous studies (Wei et al., 2022; Wang et al., 2022) defined task clusters in various ways, and we follow the T0 (Sanh et al., 2022) task taxonomy. T0 leverages the templates source software application (Bach et al., 2022) to collect English templates, which are subsequently used to form a Public Pool of Prompts (P3) for learning. We employ P3 for English meta-dataset which comprises 12 tasks and 62 datasets.

2.2 Cross-lingual Task Generalization in Instruction Tuning

Previous studies (Wang et al., 2022; Muennighoff et al., 2022) have contributed to the understanding of cross-lingual zero-shot generalization within the instruction tuning. Wang et al. (2022) first extend the boundaries by introducing both English and multilingual models trained on instruction formatted datasets. Their study construct a meta-dataset encompassing 76 task types and 1616 datasets, which included 576 datasets across 54 non-English languages. Muennighoff et al. (2022) investigate the efficacy of English-only instruction tuning in enhancing performance on non-English held-out tasks. Moreover, they introduce meta-datasets xP3 and xP3_mt, enriched with multilingual datasets

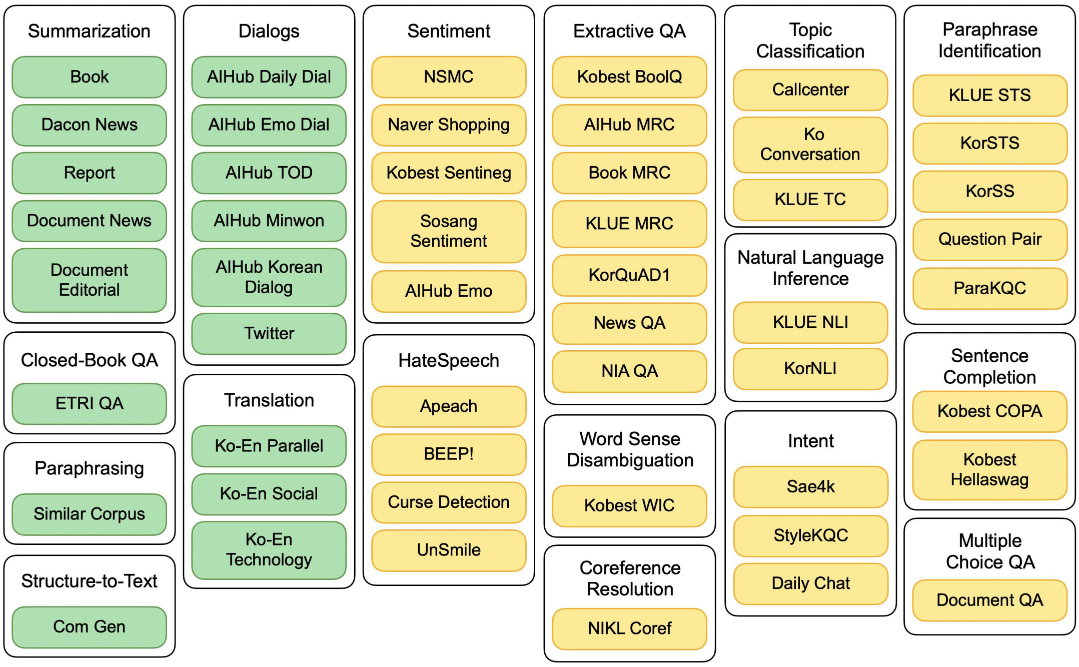

Figure 1: KORANI datasets and task taxonomy. Green datasets are NLG datasets. Yellow datasets are NLU datasets. We follow task categorization from Sanh et al. (2022)

and machine-translated templates, demonstrating further improved zero-shot performance in both English and non-English tasks.

However, the studies by Wang et al. (2022); Muennighoff et al. (2022) have some limitations. First, the majority of the training data is composed of English, with only a minor portion in non-English languages. This setup confirms the transfer effectiveness from English to other languages but fails to thoroughly explore the transfer capabilities from non-English languages to other languages (Phang et al., 2020; Chalkidis et al., 2021; Vu et al., 2022; Ouyang et al., 2022). Second, although the evaluations include a variety of non-English languages, the number of tasks per language is limited, hindering a comprehensive validation across diverse task types. Most importantly, while these studies verify the performance enhancements due to cross-lingual transfer, they do not address how these improvements compare to those achieved through monolingual instruction tuning.

To address this gap, we conduct a more comprehensive study to investigate cross-lingual zero-shot generalization in instruction tuning. Specifically, we construct and learn meta-datasets in both English and Korean and demonstrate the cross-lingual zero-shot generalization efficacy by directly comparing the performance of instruction tuning trained in other languages to that in the same language.

3 Measuring Cross-lingual Zero-shot Generalization

To investigate the effect of cross-lingual zero-shot generalization in instruction tuning, we perform instruction tuning independently for both English and Korean. We then measure the cross-lingual generalization by evaluating the models' performance on unseen tasks in the other language, respectively.

3.1 Dataset for Instruction Tuning

3.1.1 KORANI: KOReAn Natural Instructions

For the Korean instruction tuning, we introduce a novel meta-dataset named KORANI. KORANI is the first collection of various Korean NLP tasks available in the Korean research community, which then transformed into an instructional format that describes the task in plain language. The significance of our research lies in the fact that, unlike previous studies (Muennighoff et al., 2022; Li et al., 2023) that relied on machine translation, we curate and generate high-quality datasets through meticulous human effort by experts. The collection process comprises benchmark collection, instruction creation, and quality control.

Benchmark Collection Creating an instruction tuning dataset with numerous different tasks from scratch can be a resource-intensive process. To overcome this challenge, we collected 51 existing

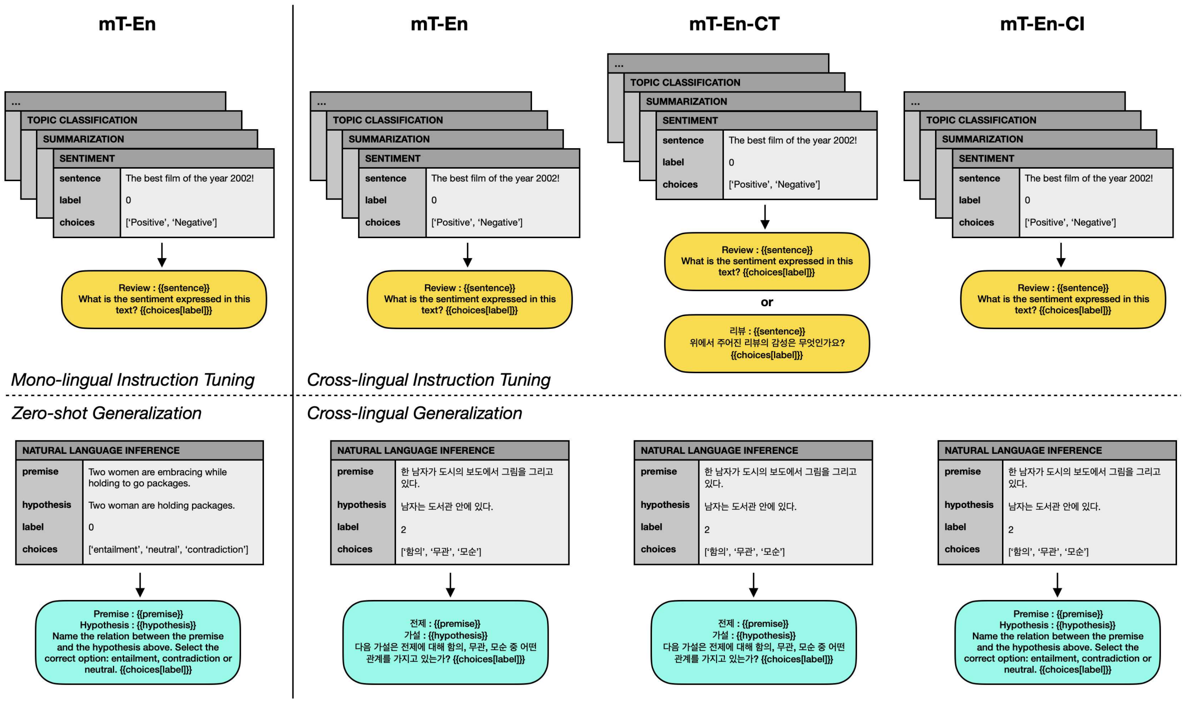
mT-En
...
TOPIC CLASSIFICATION
SUMMARIZATION
SENTIMENT
sentence
The best film of the year 2002!
label
0
choices
['Positive', 'Negative']
Review : {{sentence}}
What is the sentiment expressed in this text? {{choices[label]}}
Mono-lingual Instruction Tuning
Zero-shot Generalization
NATURAL LANGUAGE INFERENCE
premise
Two women are embracing while holding to go packages.
hypothesis
Two woman are holding packages.
label
0
choices
['entailment', 'neutral', 'contradiction']
Premise : {{premise}}
Hypothesis : {{hypothesis}}
Name the relation between the premise and the hypothesis above. Select the correct option: entailment, contradiction or neutral. {{choices[label]}}
mT-En
Cross-lingual Instruction Tuning
Cross-lingual Generalization
NATURAL LANGUAGE INFERENCE
premise
한 남자가 도시의 보도에서 그림을 그리고 있다.
hypothesis
남자는 도서관 안에 있다.
label
2
choices
['함의', '무관', '모순']
전제 : {{premise}}
가설 : {{hypothesis}}
다음 가설은 전제에 대해 함의, 무관, 모순 중 어떤 관계를 가지고 있는가? {{choices[label]}}
mT-En-CT
Review : {{sentence}}
What is the sentiment expressed in this text? {{choices[label]}}
or
리뷰 : {{sentence}}
위에서 주어진 리뷰의 감성은 무엇인가요? {{choices[label]}}
mT-En-CI
Premise : {{premise}}
Hypothesis : {{hypothesis}}
Name the relation between the premise and the hypothesis above. Select the correct option: entailment, contradiction or neutral. {{choices[label]}}

Figure 2: Comparison of model variants mT-En, mT-En-CT, and mT-En-CI on samples from Rotten Tomatoes, esNLI for P3 (Sanh et al., 2022), and KLUE NLI for KORANI. The dashed line differentiates training and evaluation, while the solid line distinguishes monolingual and cross-lingual generalization. mT-En-CT pairs English datasets with either English or Korean templates during training, and mT-En-CI pairs Korean datasets with English templates during evaluation.

Korean benchmarks from various eminent sources such as AIHub[2], Korpora[3], Github, Huggingface, KLUE[4] (Park et al., 2021), Korquad[5], and ETRI[6] including both language understanding and language generation tasks. Then, collected datasets are refined and categorized into task clusters. Some of KORANI datasets had no explicitly defined task. For those datasets, we define task and transform the dataset into an organized form based on careful consideration of the dataset's purpose and available data-labels included in the dataset. KORANI consists of 17 task clusters using heuristic rules proposed by Sanh et al. (2022) as illustrated in Figure 1. Please refer to Appendix A.2 for more details about dataset collection.

Instruction Creation For each benchmark dataset, we manually manufacture 10 natural language instructions. We collaborated with 10 experienced NLP experts to create qualitative templates. Contributors were provided a detail guide to ensure that they utilize a various data-labels in the dataset. To create diverse and well-refined templates, we encouraged contributors to be open to their own style while providing strict guidelines for grammatical accuracy and clarity in natural language instructions.

Quality Control For quality control, we removed duplicate instances, and adjusted label imbalances in each task. We went through a peer-review session with two other expert contributors per dataset five times at minimum to ensure the quality of templates. The extensive cross-validation was iteratively performed until reviewers were content with the quality of the datasets.

3.1.2 English Instruction Tuning Benchmarks

We utilized a Public Pool of Prompts (P3) following Sanh et al. (2022) as the English meta-dataset. Our experiment involves 62 datasets and templates for each task provided by Bach et al. (2022). We split the datasets into 12 task clusters shown in Appendix A.3.

3.1.3 Statistics of KORANI and P3

Table 1 shows various statistics and comparisons between the KORANI and P3. KORANI consists of 51 tasks divided into 17 clusters, each comprising

[2] https://www.aihub.or.kr
[3] https://ko-nlp.github.io/Korpora/en-docs
[4] https://klue-benchmark.com
[5] https://korquad.github.io
[6] https://nanum.etri.re.kr

Statistics	KORANI	P3
# of datasets	51	62
# of NLU datasets	34	51
# of NLG datasets	17	11
# of tasks	17	12
avg. # of templates (per dataset)	10	8.02
avg. # of cross-lingual templates (per dataset)	3.76	3.45
avg. # of instances (per dataset)	4,768	4,388
avg. # of input tokens (per dataset)	179.8	187.64
avg. # of output tokens (per dataset)	16.35	11.06

Table 1: Statistics of KORANI and P3. Training instances per dataset are limited to 5k maximum.

a comparable number of Natural Language Generation (NLG) and Natural Language Understanding (NLU) tasks. In contrast, the P3 is more heavily focused on NLU tasks.

3.2 Addressing Templates Misalignment Challenges in Cross-Lingual Instruction Tuning Scenarios

To evaluate the cross-lingual zero-shot generalization, we conduct instruction tuning in one language and assess the model's performance on unseen tasks in the other language. However, since the training templates and inference templates are taken from different language datasets (KORANI and P3), it raises the possibility of performance degradation from template misalignment, which might lead to suboptimal performance (Sanh et al., 2022; Wang et al., 2022; Muennighoff et al., 2022; Liang et al., 2023; Sun et al., 2024). The misalignment primarily occurs in two aspects. One is the linguistic misalignment of the templates (Muennighoff et al., 2022), which stems from the grammatical and semantic differences between the two languages. The other is the misalignment in the instructional format (Kung and Peng, 2023; Yin et al., 2023; Liang et al., 2023; Sun et al., 2024), which stems from template style differences such as ordering description in templates and level of explanation detail about the task.

We introduce cross-lingual templates to mitigate the challenges of template misalignment in cross-lingual instruction tuning scenarios. To maximize alignment between training and inference templates, we align the language and instructional format of the templates similar to targeting evaluation tasks. We create an average of 3.76 and 3.45 cross-lingual templates each for KORANI and P3 meta datasets as shown in Table 1. See the Appendix A.4 for more details on the creation process of cross-lingual templates.

We propose two approaches to integrate cross-lingual templates with cross-lingual settings. The first approach utilizes cross-lingual templates during the training phase, while the second approach employs cross-lingual templates during the inference phase to align language and instructional format in training and inference. Both strategies are meticulously designed to uphold structural similarity between templates used during training and inference. The linguistic and instructional format alignment empower the model to effectively adapt when it is presented with a new instruction for an unseen task.

3.3 Model

To assess the zero-shot generalization capability of instruction tuning and its cross-lingual transferability between Korean and English, we employ mT5 (Xue et al., 2021) models as the core model. The mT5 model is a publicly available multilingual model trained in 101 languages, including both English and Korean. The mT5 models encompass a range of sizes, from 300M to 13B parameters, and we employ 1.3B to 13B for our experiments.

We assess cross-lingual generalization by instruction tuning in one language and evaluating unseen tasks in the other language for both English and Korean. Moreover, we introduce cross-lingual templates in the training or inference phases to investigate the advantage of instruction alignment. For this scenario, we introduce the following model variants:

- **mT-Ko, mT-En:** Models trained on KORANI and P3 datasets, respectively.
- **mT-Ko-CT, mT-En-CT**: Models trained on KORANI and P3 datasets, respectively by incorporating cross-lingual templates during training only, and inferenced with original templates.
- **mT-Ko-CI, mT-En-CI**: Models trained on KORANI and P3 datasets respectively with original templates only, and inferenced with cross-lingual templates.

The postfix **CT** and **CI** denote **C**ross-lingual instruction **T**raining, and **C**ross-lingual instruction **I**nference respectively.

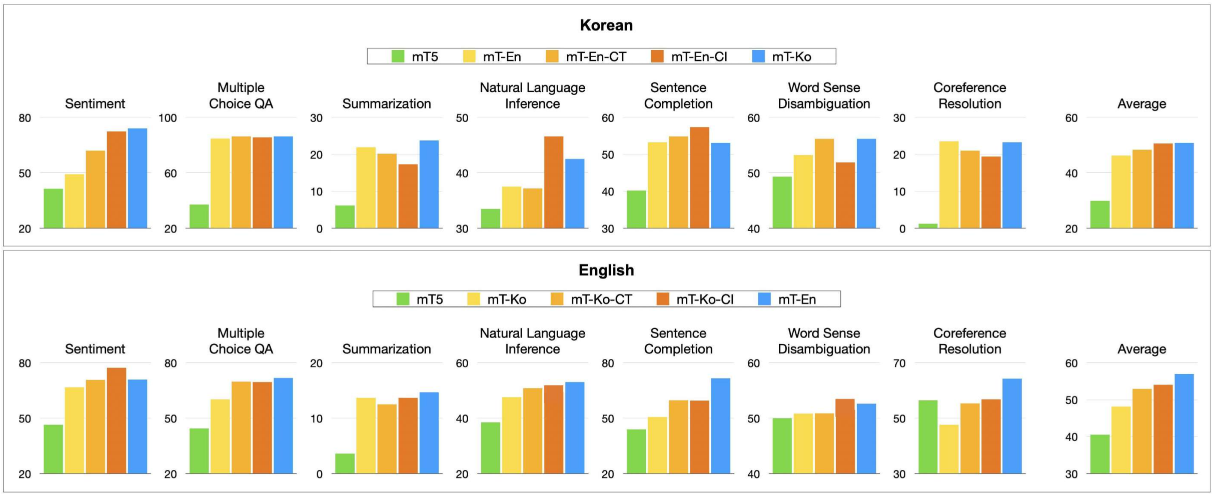

Figure 3: Performance of zero-shot and cross-lingual generalization. Scores are datasets average for each task cluster. The first row denotes KORANI unseen tasks, and the second row denotes P3 unseen tasks. Average chart averages seven different task results. Appendix E.1 breaks down the performance by datasets.

4 Experimental Setup

4.1 Training

To evaluate the zero-shot performance for various tasks, we set held-out tasks following previous studies (Sanh et al., 2022; Muennighoff et al., 2022). In this scheme, when training a meta-dataset in a different language, we deliberately exclude tasks that correspond to the pre-defined held-out tasks in the meta-dataset. We have devised two distinct held-out settings. The first group, following Sanh et al. (2022), encompasses four tasks: natural language inference, sentence completion, coreference resolution, and word sense disambiguation. In addition to the existing settings, a second group is formed to further validate the trend of cross-lingual generalization for more tasks. The second group consists of three tasks: sentiment analysis, summarization, and multiple-choice QA.

During the training, we employ 10 distinctive templates for each dataset. For CT models, we partially replace original templates with cross-lingual templates.

We configure validation from training datasets and select the model that showed the best performance in the validation. Our experiment is in a true zero-shot setting, as we do not use any examples from held-out tasks for checkpoint selection.

We also limit the number of examples in each dataset to 5k to avoid a skewed distribution between tasks, which end up around 180k instances per train. For more information on training, see Appendix B.1.

4.2 Evaluation

We randomly sample three templates and measure the average score for each dataset. For the classification task, we employ rank classification (Brown et al., 2020) and for the generation task, we report ROUGE-L (Lin, 2004) score for model performance, following previous work (Wang et al., 2022).

The CI models follow a consistent training approach as the base models (mT-Ko and mT-En) but have a key distinction in their evaluation process by utilizing cross-lingual templates. Therefore, unlike all other models, including the CT model, the CI model gauges its performance across three randomly sampled cross-lingual templates different from the original templates.

5 Result

5.1 Cross-lingual Transfer between Korean and English

To assess the impact of cross-lingual zero-shot generalization, we initially conduct instruction tuning in one language and evaluate the model's performance on unseen tasks in the other language. As depicted in Figure 3, both languages exhibit notable performance enhancements even when subjected to instruction tuning conducted in a different language. Moreover, in certain tasks, the achieved performance closely resembles that of the model trained within the same language. Specifically, tasks like multiple-choice QA, summarization, and sentence completion of Korean evaluation display comparable performance between mT-En

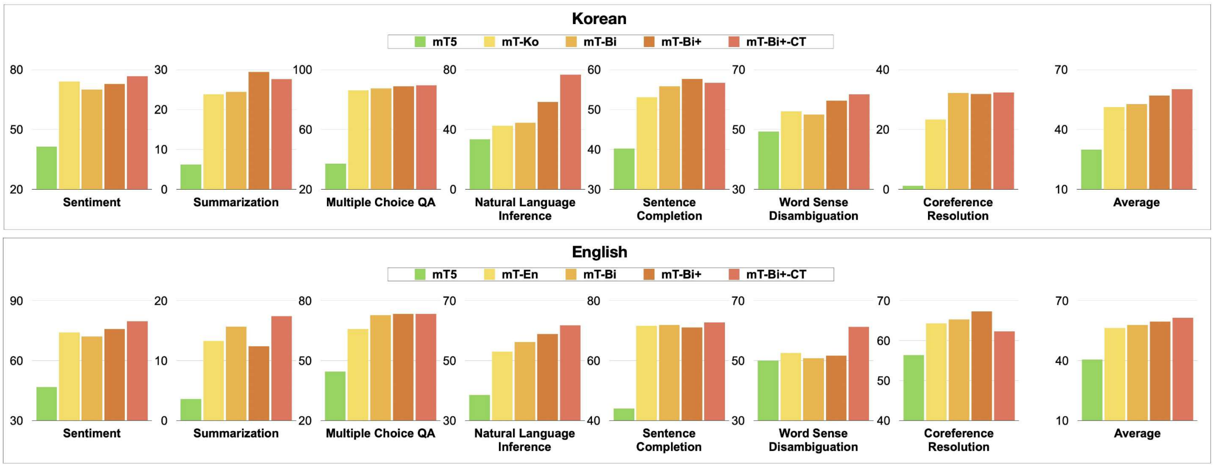

Figure 4: Bilingual instruction tuning performance in KORANI and P3. mT-Bi+-CT employs the CT training method for non-target language datasets only. Appendix D covers additional experiments on the cross-lingual template, and E.2 breaks down the performance by datasets.

and the model mT-Ko. For English evaluation, the mT-Ko demonstrates performance akin to that of mT-En in sentiment analysis and summarization tasks.

5.2 Effect of Cross-lingual Templates in Cross-lingual Generalization

Furthermore, CT and CI models that incorporate well-aligned cross-lingual templates show notable performance improvements across most tasks in both languages. Specifically, in the Korean evaluation, mT-En-CT and mT-En-CI outperform mT-En. Similarly, in the English evaluation, mT-Ko-CT and mT-Ko-CI surpass mT-Ko. This finding highlights the importance of well-aligned templates in facilitating effective cross-lingual generalization.

The key takeaway from our experiments is that the mT-En-CI and mT-Ko-CI models, trained in a different language, achieve performance comparable to the mT-Ko and mT-En models, which are trained in the same language. This consistent trend is observed across the majority of tasks during evaluations in both Korean and English. Notably, the CI models outperform, particularly excelling in certain tasks like sentiment analysis and sentence completion. Specifically, models trained on the KORANI dataset display robust performance in sentiment analysis, while the CI models trained on the P3 dataset demonstrate exceptional sentence completion capabilities, regardless of the language evaluated. These findings indicate that training on relevant tasks in a different language can still yield significant performance, potentially even surpassing that of monolingual instruction tuning. This underscores that focusing on relevant tasks is more important than adhering to linguistic congruence for unseen tasks. For a more detailed analysis, please refer to the Appendix C.

6 Further Analysis

6.1 Bilingual Instruction Tuning

Instruction tuning in a single language sufficiently shows cross-lingual zero-shot generalization. To delve deeper into the potential synergistic effects arising from the utilization of two languages in instruction tuning together, we introduce bilingual instruction tuning. This approach combines and jointly trains two meta-datasets, KORANI and P3, then compares the performance of unseen tasks with single language instruction tuning. This experiment involves the following additional model variants:

- **mT-Bi** models are trained on both the KORANI and P3 datasets using bilingual instruction tuning.
- **mT-Bi+** maintains the settings of mT-Bi but includes held-out tasks from a non-evaluating language.
- **mT-Bi+-CT** employs the same dataset composition as mT-Bi+ and further incorporates cross-lingual templates from the non-evaluating language's meta-dataset during the training phase.

Figure 4 shows that training the model on a mixture of two meta-datasets results in improved performance compared to single meta-dataset training for both languages. We speculate that similar to the trend in monolingual instruction tuning (Wei et al., 2022; Sanh et al., 2022), where increased task diversity enhances performance, learning a broader range of tasks irrespective of language has also led to performance improvements in cross-lingual instruction tuning.

Furthermore, mT-Bi+, the model trained by adding datasets corresponding to held-out tasks in the other language, demonstrates improved performance. This result underscores that incorporating aligned datasets can guide the model to more explicitly learn the targeted unseen tasks, thereby enhancing cross-task generalization.

Lastly, when considering the mT-Bi+-CT model, which integrates cross-lingual templates into the mT-Bi+ during the training phase, consistent performance enhancements are observed for both languages. This trend aligns with Section 5.1, emphasizing alignment of instructions facilitates improved adaptation of the model to unseen tasks.

6.2 Template alignment: Linguistic Or Instructional Format

In this section, we analyze whether the performance improvements in cross-lingual transfer through instruction alignment originate from linguistic factors or formatting. To do this, we employ mT-Ko and two of its variants and evaluate their performance on the held-out tasks from the P3 benchmark. The first variant, mT-Ko-Trans, merely employs a translated version of P3 templates into Korean during the inference phase. This variant aligns linguistic aspects of templates between training and inference. The second variant, mT-Ko-CI, as mentioned in Section 3.3, encompasses instruction alignment that considers both linguistic and instructional format aspects during inference. Further illustrative examples are available in the Appendix A.5.

Table 2 shows that mT-Ko-Trans demonstrates consistently improved performance compared to the mT-Ko evaluated using P3 templates, across most of the tasks. This observation proves that notable performance enhancement is achievable through linguistic alignment alone, as it guides the model to better adapt to new unseen templates. When comparing the performance of mT-Ko-Trans with that of mT-Ko-CI, it becomes evident that the latter achieves higher performance. This result is attributed to the fact that while both approaches entail linguistic alignment between the training and evaluation templates, mT-Ko-CI additionally gains cross-task generalization through the alignment of structural formatting between training and evaluation templates. The result highlights that both linguistic and instructional format alignment is important in cross-lingual generalization.

	NLI	SC	WSC	CR	SENT	SUM	MUL	AVG
mT-Ko	47.6	50.6	50.8	47.6	66.0	**13.7**	60.1	48.1
mT-Ko-Trans	51.4	54	52.3	48.9	74.1	13.3	65.8	51.4
mT-Ko-CI	**52.2**	**55.3**	**53.3**	**56.8**	**77.0**	**13.7**	**70.5**	**54.1**

Table 2: Performance of held-out P3 datasets with mT-Ko-Trans model *(linguistic alignment only)*, and mT-Ko-CI model *(linguistic and instructional format alignment)*. The best comparable performances are **bolded**. Details are on Appendix E.3

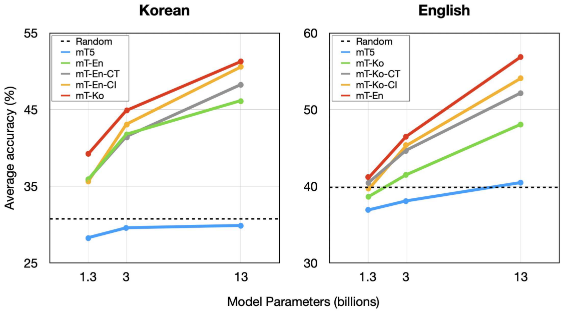

Figure 5: Model performance vs. size. The random line represents the average score random choice in the options list for classification tasks, and the ROUGE-L score of a copy of input for generation tasks. Appendix E.4 breaks down the performance by datasets.

6.3 Scaling Laws

In our final ablation experiment, we investigate how the cross-lingual generalization of instruction tuning evolves with model size. Using the same model variants and cluster split as in Section 5, we assess cross-lingual generalization performance across model sizes of 1.3B, 3B, and 13B.

Figure 5 illustrates the average performance of unseen tasks in both Korean and English. As model size increases, we observe performance improvements for all instruction-tuned models. Additionally, across various model sizes, models that incorporate cross-lingual templates exhibit higher performance, similar to the trends in the results of Section 5. Particularly, for Korean evaluation, when the model size reaches 13B, mT-En-CI achieves comparable average performance to mT-Ko. In contrast, for English tasks, as the model size increases, the performance of mT-Ko-CI improves, but mT-En still exhibits performance differences. We conjecture that this trend may be attributed to the predominance of NLU tasks within the held-out task set. Given the profusion of NLU tasks within P3 compared to KORANI, mT-En, trained on P3, may possess an advantage in comprehending these held-out tasks in Korean and English. Detailed analyses of task-specific performance changes concerning model size are provided in Appendix E.4.

7 Conclusion

Our research contributes to a deep understanding of the cross-lingual zero-shot generalization effect and its benefits by leveraging the novel KORANI meta-dataset to compare cross-lingual and monolingual instruction tuning directly. The experimental results indicate that cross-lingual instruction tuning can match or even exceed the performance of monolingual tuning. Our findings highlight the importance of training relevant data across diverse languages rather than strictly maintaining linguistic consistency in unseen tasks. The successful application of cross-lingual templates, which ensure consistency in both language and format, further validates the potential of cross-lingual instruction tuning. These discoveries present cross-lingual instruction tuning not just as an auxiliary strategy but as a potential alternative to monolingual methods, especially in low-resource language scenarios.

Limitations

In this work, we primarily examine cross-lingual instruction tuning between Korean and English, which, while informative, provides a partial view of cross-lingual generalization due to the exclusion of other languages. Moreover, we utilized the mT5 model, a multilingual model trained in various languages, but more focused on English than Korean. Lastly, a difference in task cluster distribution between KORANI and P3 makes the result vague since the composition of datasets holds a significant effect.

Acknowledgements

We would like to express our heartfelt gratitude to Professor Moontae Lee for his invaluable guidance, discussions, and feedback throughout the course of this research. We also extend our sincere thanks to Joel Jang and Seonghyeon Ye for their thoughtful feedback on the paper.

References

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6974–6996. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, On-*

line, July 5-10, 2020, pages 8440–8451. Association for Computational Linguistics.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 14702–14729. PMLR.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Shihao Liang, Kunlun Zhu, Runchu Tian, Yujia Qin, Huadong Wang, Xin Cong, Zhiyuan Liu, Xiaojiang Liu, and Maosong Sun. 2023. Exploring format consistency for instruction tuning. *CoRR*, abs/2307.15504.

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.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. *CoRR*, abs/2211.02001.

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. 2022. Crosslingual generalization through multitask finetuning.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 557–575. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien

Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *The Twelfth International Conference on Learning Representations*.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9279–9300. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2856–2878. Association for Computational Linguistics.

A Dataset

A.1 Public Release

To foster active research on instruction tuning in the Korean community, we have made CSV files available on GitHub, containing instruction-templated inputs and outputs for benchmarks that have free copyright of derivative works. For the five out of 51 datasets that have limited copyright of derivative works, we have provided a method to download the data and preprocess code with underlying templates. Our objective is to enable researchers to readily utilize these datasets for instruction tuning research while respecting copyright laws[7].

A.2 Example of Raw Data Transformations

Open-source datasets included in KORANI are typically characterized by predefined tasks and various labels. We utilize these labels, or generate new ones, to create templates. Even if the original purpose of the dataset differs, we leverage the labels to develop turn-around tasks, similar to approaches used in Sanh et al. (2022) and Wei et al. (2022). For instance, in the case of dialog tasks, such as AIHub TOD, the presence of labels indicating the topic of conversation enabled us to create topic classification tasks. For summarization tasks, we create labels by extracting keywords using KeyBERT and classify topics of the document using the TF-IDF

[7] For detailed information about data sources and license information, please refer to the following link: `https://github.com/CHLee0801/KORANI-Instruction-Tuning`.

algorithm. We then utilize these labels to enrich the instructions, thereby incorporating contextual information about the conversation's topic and enhancing the overall comprehensiveness of the instructions.

A.3 P3 Datasets and Task Taxonomy

Please refer to Figure 6.

A.4 Cross-lingual Templates Generation

To encompass linguistic attributes for cross-lingual templates, instead of just translating corresponding task templates from another language, a concerted effort is exerted to extract the underlying semantics and salient components inherent to the templates. Subsequently, these components are meticulously organized and structured in a manner that not only accentuates the intrinsic qualities of the target language but also ensures the retention of task-specific characteristics. Within the context of generating cross-lingual templates, a fundamental aspect involves the meticulous alignment of structural elements across languages. Furthermore, the configuration of choices for classification is adapted to the auxiliary language, unless the task inherently demands distinct sentence or phrase-based choices for each instance—illustratively observed in tasks like sentence completion and multiple-choice QA. The strategic deployment of demonstrative pronouns and the method of incorporating meta-data are thoughtfully tailored by the structural framework of the templates in question, thereby underlining the significance of syntactic considerations. Please refer to Section for specific examples.

A.5 Illustrative examples of cross-lingual templates and translated templates

The examples below demonstrate the original English templates from Promptsource, cross-lingual templates for mT-Ko-Trans, and mT-Ko-CI. mT-Ko-Trans is a translated version of English templates, and the structural format is identical as well, while mT-Ko-CI is more aligned with templates in KORANI datasets. Table 3, Table 4, and Table 5 illustrate the examples of cross-lingual templates from P3 datasets. Moreover, Table 6, Table 7, and Table 8 illustrate the examples of cross-lingual templates from KORANI datasets. All meta-data of the dataset are represented in double brackets.

<table>
<tr><th></th><th>P3 Templates</th></tr>
<tr><td><b>Input</b></td><td>{{document}}\n\n ===\n\nWrite a summary of the text above :</td></tr>
<tr><td><b>Output</b></td><td>{{summary}}</td></tr>
<tr><td></td><td>Translated Templates</td></tr>
<tr><td><b>Input</b></td><td>{{document}}\n\n ===\n\n위 글을 영어로 요약하시오.</td></tr>
<tr><td><b>Output</b></td><td>{{summary}}</td></tr>
<tr><td></td><td>Cross-lingual Templates</td></tr>
<tr><td><b>Input</b></td><td>다음은 글을 읽고 요약하는 문제입니다.\n {{document}}\n위 글을 영어로 요약하세요.</td></tr>
<tr><td><b>Output</b></td><td>{{summary}}</td></tr>
</table>

Table 3: Instruction examples of XSum : Summarization.

<table>
<tr><th></th><th>P3 Templates</th></tr>
<tr><td><b>Input</b></td><td>{{text}} In the previous sentence, does the pronoun "{{span2_text}}" refer to {{span1_text}}? Yes or no?</td></tr>
<tr><td><b>Choices</b></td><td>Yes ||| no</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
<tr><td></td><td>Translated Templates</td></tr>
<tr><td><b>Input</b></td><td>{{text}} 이전 문장에서 단어 "{{span1_text}}" 는 "{{span2_text}}"를 참조하는가? 예, 아니오</td></tr>
<tr><td><b>Choices</b></td><td>예 ||| 아니오</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
<tr><td></td><td>Cross-lingual Templates</td></tr>
<tr><td><b>Input</b></td><td>글에서 같은 것을 의미하는 다른 두 단어는 서로를 참조하는 관계이다. 문장: {{text}}\n위 문장에서 단어 {{span1_text}}와 "{{span2_text}}" 의 뜻이 같은가?</td></tr>
<tr><td><b>Choices</b></td><td>예 ||| 아니오</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
</table>

Table 4: Instructions examples of WSC : Coreference Resolution.

<table>
<tr><th></th><th>P3 Templates</th></tr>
<tr><td><b>Input</b></td><td>{{text}}. What is the emotion expressed in this message?</td></tr>
<tr><td><b>Choices</b></td><td>sadness ||| joy ||| love ||| anger ||| fear ||| surprise</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
<tr><td></td><td>Translated Templates</td></tr>
<tr><td><b>Input</b></td><td>{{text}} 이 메세지에 나타난 감정은 무엇인가?</td></tr>
<tr><td><b>Choices</b></td><td>슬픔 ||| 기쁨 ||| 사랑 ||| 화남 ||| 공포 ||| 놀람</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
<tr><td></td><td>Cross-lingual Templates</td></tr>
<tr><td><b>Input</b></td><td>감정 분류 태스크이다.\n{{text}}\n 위에서 확인할 수 있는 사람의 감정을 알려줘.</td></tr>
<tr><td><b>Choices</b></td><td>슬픔 ||| 기쁨 ||| 사랑 ||| 화남 ||| 공포 ||| 놀람</td></tr>
<tr><td><b>Output</b></td><td>{{ answer_choices [label] }}</td></tr>
</table>

Table 5: Instruction examples of Emotion : Sentiment.

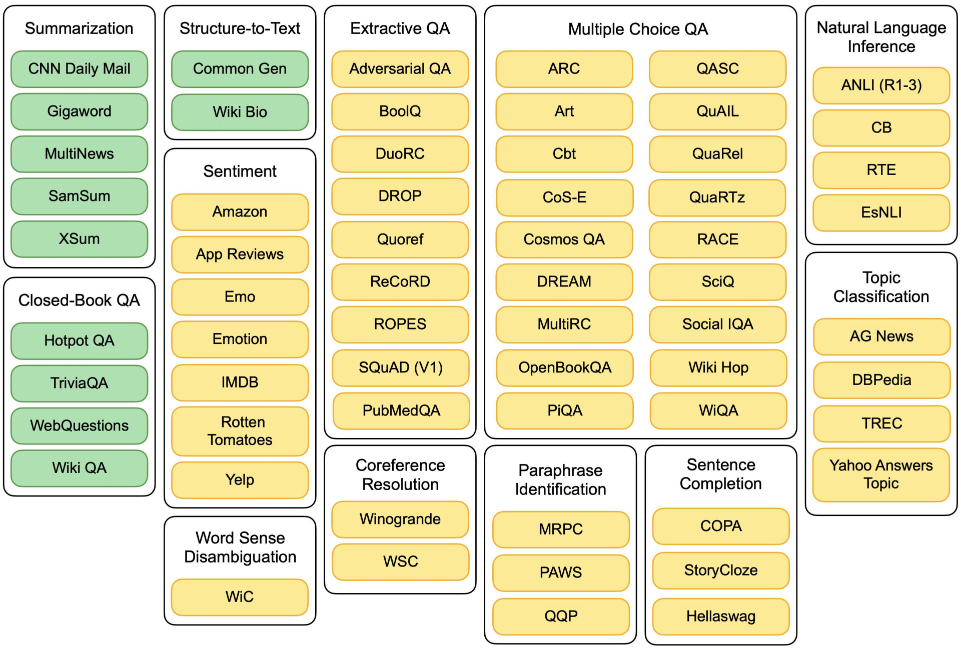

Figure 6: P3 datasets and task taxonomy. Green datasets are NLG datasets, Yellow datasets are NLU datasets. We follow task categorization from Sanh et al. (2022)

	KORANI Templates
Input	다음은 보기가 주어져서 정답을 고르는 기계독해 문제이다. 답을 {{choices}} 중에서 고르시오.\n {{context}}\n{{question}}
Choices	choices[0] ||| choices[1] ||| choices[2] ||| choices[3]
Output	{{ answer_choices [label] }}
	Cross-lingual Templates
Input	Read the following context and choose the best option to answer the question. \nContext: {{context}}\nQuestion: {{question}}\nOptions:\n {{choices}}
Choices	choices[0] ||| choices[1] ||| choices[2] ||| choices[3]
Output	{{ answer_choices [label] }}

Table 6: Instruction examples of Document QA : Multiple Choice QA.

	KORANI Templates
Input	자연어 추론 문제이다. 이 문제는 전제가 참이라고 가정할 때, 가설의 내용이 참(함의)인지, 거짓(모순)인지, 혹은 알 수 없는지(무관)에 따라 관계가 분류된다. 전제와 가설의 관계를 유추하라.\n전제: {{premise}}\n 가설: {{hypothesis}}\n선택지: {{choices}}
Choices	함의 ||| 모순 ||| 무관
Output	{{ answer_choices [label] }}
	Cross-lingual Templates
Input	{{premise}} Using only the above description and what you know about the world, ""{{hypothesis}}"" is definitely correct, incorrect, or inconclusive?
Choices	Correct ||| Inconclusive ||| Incorrect
Output	{{ answer_choices [label] }}

Table 7: Instruction examples of KLUE NLI : Natural Language Inference.

	KORANI Templates
Input	아래 두 문장에서 [{{word}}]의 뜻이 같은지 판별하시오.\n{{sentence1}}\n{{sentence2}} \n선택지: {{choices}}
Choices	예 ||| 아니오
Output	{{ answer_choices [label] }}
	Cross-lingual Templates
Input	Does the word [{{word}}] have the same meaning in these two sentences?\n {{sentence1}}\n{{sentence2}}\n{{choices}}
Choices	Yes ||| No
Output	{{ answer_choices [label] }}

Table 8: Instruction examples of Kobest WiC : Word Sense Disambiguation.

B Training and evalauation

B.1 training

The second group consists of three tasks: sentiment analysis, summarization, and multiple-choice QA. We choose sentiment analysis as a held-out task because it has the potential to discern whether the model effectively comprehends semantic nuances across languages. We also hold out summarization to verify the extent of task generalization within generative tasks. Lastly, the decision to hold out multiple-choice QA stems from the intricate nature of the task's choices, which demand a nuanced understanding of linguistic subtleties beyond mere structural or template patterns.

We truncate input and target sequences to 768

Models	NLI	SC	WSD	CR
mT5	33.5	40.2	49.3	1.2
mT-Ko	42.5	53.1	56.1	23.3
mT-Bi	44.5	55.8	55.0	32.2
mT-Bi-CT	46.8	**58.2**	58.0	31.1
mT-Bi+	58.4	57.7	59.6	31.9
mT-Bi+-CT	**76.7**	56.7	**61.8**	**32.4**

Table 9: Performance of Bilingual Instruction Tuning on KORANI Evaluation Benchmarks (Natural Language Inference, Sentence Completion, Word Sense Disambiguation, Coreference Resolution) for Different Models.

and 256 tokens, respectively. We train all models with a batch size of 64 using AdamW Optimizer with a learning rate of 1e-5. We also train all models for 1 epoch and save checkpoints for every 600 steps to select checkpoints for evaluation.

For validation, we sample 100 examples from the validation splits of each training dataset. We measure the performance of each dataset and aggregate them to perform checkpoint selection. This approach avails our experiment in a true zero-shot setting, as we do not use any examples from held-out tasks for checkpoint selection.

C More Analysis on Cross-Lingual Instruction Tuning

We claim the performance enhancement stems from the inclusion of relevant tasks in the training phase, with the enhancement increasing when the language is aligned. For example, we hypothesize that including the "piqa" dataset in mT-En training boosts performance on sentence completion tasks, as "piqa", a multiple-choice QA dataset, closely resembles sentence completion tasks. Furthermore, for sentiment analysis, mT-Ko-CT and mT-Ko-CI demonstrate superior performance in English evaluations, likely due to the inclusion of the hatespeech task from the KORANI datasets. These datasets, which are designed to identify sentence toxicity, may align well with sentiment analysis tasks.

D Additional Experiment on Bilingual Instruction Tuning

We conducted an extra experiment on mT-Bi (CT) for specific tasks on Table 9. The trend observed aligns with the findings presented in Figure 4.

E Results Breakdown

This section shows the full results for all datasets we evaluate. We show the performance of the models using randomly chosen three templates per dataset with the best performance on the dev set. All results are average scores of three templates. For efficient evaluation, we sample a maximum of 1000 instances for all generative datasets. These include summarization datasets for KORANI and P3, and coreference resolution datasets for KORANI. We use greedy search for all generative tasks.

E.1 Zero-shot cross-lingual generalization performance breakdown

Table 10 and Table 11 break down the scores of zero-shot cross-lingual generalization performance of KORANI and P3 respectively.

E.2 Bilingual Instruction Tuning Performance Breakdown

Table 12 and Table 13 break down the scores of bilingual instruction tuning performance of KORANI and P3 respectively.

E.3 Template Alignment Performance Breakdown

Table 14 breaks down the scores of bilingual instruction tuning performance of KORANI and P3 respectively.

E.4 Scale up.

Figure 7 breaks down the scores average of each task by scaling up from 1.3b to 13b.

Model	Sentiment					Avg.	Summarization					Avg.	Mul. QA
	AIHub Emo	Kobest Sentiment	NSMC	Naver Shopping	Sosang Sentiment		Book	Dacon News	Doc. Edi.	Doc. News	Report		Doc. QA
mT5	18.2	50.4	54.8	50	33.5	41.4	5.8	6.8	5.6	8	4.8	6.2	37.4
mT-En	53.5	54.5	51.9	51	35.3	49.2	26.3	25.1	15.8	23.3	**19.2**	21.9	84.7
mT-En-CT	**55**	83.4	63.4	62.1	46.2	62	23.4	23.6	15.3	20.8	17.9	20.2	**86.4**
mT-En-CI	46.7	90.7	**81.6**	**87.2**	**55.8**	72.4	17.6	20.8	13.8	20.1	14.3	17.3	85.6
mT-Ko	54.4	**96.4**	80.2	84.1	55.5	**74.1**	**30**	**28.6**	**17.5**	**24.4**	18.6	**23.8**	86.2

Model	NLI		Avg.	Sent. Comp.		Avg.	Coref. Resol	WSD	Total Avg.
	KLUE NLI	KorNLI		Kobest Copa	Kobest Hellas.		NIKL Coref	Kobest WiC	
mT5	33.4	33.6	33.5	53.3	27	40.2	1.2	49.3	29.9
mT-En	39.7	35.3	37.5	66.9	39.7	53.3	**23.5**	53.2	46.2
mT-En-CT	39.4	34.9	37.2	66.6	43.2	54.9	21	**56.1**	48.3
mT-En-CI	**58.7**	**52.1**	**55.4**	**80.8**	34	**57.4**	22	52.7	**51.8**
mT-Ko	44.7	40.3	42.5	62.1	**44**	53.1	23.3	**56.1**	51.3

Table 10: KORANI zero-shot cross-lingual generalization performance breakdown. The best comparable performances are **bolded** and second best underlined.

Model	Sentiment				Avg.	Summarization					Avg.	Multiple-Choice QA					Avg.
	Emot.	Rot. Tom.	Ama.	IMDB		Mul.News	Sam.	CNN.	XSum	Giga.		Dream	Mul.RC	PiQA	QASC	RACE	
mT5	33.4	50.4	49.7	53.5	46.8	6.4	3.1	3.8	2.4	2.4	3.6	36.6	57.2	52.4	49.2	27.3	44.5
mT-Ko	35.8	67.7	84.9	75.7	66	**7.8**	11	**15.1**	11.9	**22.9**	13.7	59.1	65.3	57.2	74.6	44.2	60.1
mT-Ko-CT	43.6	84.1	76.4	77.7	70.5	6.9	9	14.5	11.8	20.2	12.5	74.8	77.8	**62.3**	91.4	46	70.5
mT-Ko-CI	**49.4**	77.8	**91.2**	**89.6**	**77**	6.7	**20.2**	13.1	10.3	18.3	13.7	69	**79.8**	58.2	91	**54.6**	70.5
mT-En	40.2	**84.5**	80	77.7	70.6	7	17.1	14.3	**12.7**	22.4	**14.7**	**77.6**	78.2	61.1	**93.8**	47.7	**71.7**

Model	Natural Language Inference						Avg.	Sentence Completion			Avg.	Coref. Resol.		Avg.	WSD	Total Avg.
	RTE	CB	AN. R1	AN. R2	AN. R3	EsNLI		COPA	Hellasw.	StoryC.		Winogr.	WSC		WiC	
mT5	47.3	50.6	32.8	33.4	33	33.9	38.5	56	26	49.9	44	49.2	63.5	56.4	50	40.5
mT-Ko	65.3	75	36.2	34	37.5	37.5	47.6	63.7	28.9	59.1	50.6	51.9	43.3	47.6	50.8	48.1
mT-Ko-CT	76.8	83.3	38.2	34.6	37.4	34.3	50.8	66.3	31.8	67.6	55.2	52.5	58	55.3	50.8	52.2
mT-Ko-CI	79.5	79.2	33.6	33.8	33.0	**53.3**	52.2	63.3	**35.4**	67.3	55.3	51.6	61.9	56.8	**53.3**	54.1
mT-En	**81.3**	**85.1**	**40.5**	**36.5**	**40.2**	34.4	**53**	**85.3**	34.9	**94.7**	**71.6**	**61.6**	**67**	**64.3**	52.6	**56.9**

Table 11: P3 zero-shot cross-lingual generalization performance breakdown. The best comparable performances are **bolded** and second best underlined.

Model	Sentiment					Avg.	Summarization					Avg.	Mul. QA
	AIHub Emo	Kobest Sentiment	NSMC	Naver Shopping	Sosang Sentiment		Book	Dacon News	Doc. Edi.	Doc. News	Report		Doc. QA
mT5	18.2	50.4	54.8	50	33.5	41.4	5.8	6.8	5.6	8	4.8	6.2	37.4
mT-Ko	**54.4**	**96.4**	80.2	84.1	55.5	74.1	30	28.6	17.5	24.4	18.6	23.8	86.2
mT-Bi	52.5	96.2	76.8	72.8	51.7	70	30.9	29.5	18.7	23.9	19.1	24.4	87.4
mT-Bi+	54	96.1	80.7	78.9	54.9	72.9	**32.6**	**36.5**	**21.2**	**32.2**	**24.5**	**29.4**	88.9
mT-Bi+-CT	49.6	96	**84.5**	**92.2**	**61.2**	**76.7**	32.4	32.9	**21.2**	27.9	23.4	27.6	**89.5**

Model	NLI			Sent. Comp.			Coref. Resol	WSD	Total Avg.
	KLUE NLI	KorNLI	Avg.	Kobest Copa	Kobest Hellas.	Avg.	NIKL Coref	Kobest WiC	
mT5	33.4	33.6	33.5	53.3	27	40.2	1.2	49.3	29.9
mT-Ko	44.7	40.3	42.5	62.1	44	53.1	23.3	56.1	51.3
mT-Bi	46.6	42.4	44.5	66	45.5	55.8	32.2	55	52.6
mT-Bi+	55	61.8	58.4	**67.7**	**47.6**	**57.7**	31.9	59.6	56.5
mT-Bi+-CT	**82.7**	**70.6**	**76.7**	**67.7**	45.6	56.7	**32.4**	**61.8**	**59.7**

Table 12: KORANI bilingual instruction tuning performance breakdown. The best comparable performances are **bolded** and second best underlined.

Model	Sentiment					Summarization						Multiple-Choice QA					
	Emot.	**Rot. Tom.**	**Ama.**	**IMDB**	**Avg.**	**Mul.News**	**Sam.**	**CNN.**	**XSum**	**Giga.**	**Avg.**	**Dream**	**Mul.RC**	**PiQA**	**QASC**	**RACE**	**Avg.**
mT5	33.4	50.4	49.7	53.5	46.8	6.4	3.1	3.8	2.4	2.4	3.6	36.6	57.2	52.4	49.2	27.3	44.5
mT-En	40.2	84.5	80	77.7	70.6	7	17.1	14.3	12.7	22.4	14.7	77.6	78.2	61.1	93.8	47.7	71.7
mT-Bi	47.2	83.5	80	77.8	72.1	7.6	15.8	13.8	13.1	**28.4**	15.7	80.1	**81.2**	60.7	94.7	47.2	72.8
mT-Bi+	**54.3**	87.6	81	**80.3**	75.8	**10.9**	**25**	18.7	14.1	27.9	13.7	80.9	80.6	61.9	**95.4**	**48.1**	73.4
mT-Bi+-CT	51.8	**89.9**	**96.7**	**80.3**	**79.7**	10.7	21.5	**19**	**14.4**	27.9	**18.7**	**81.7**	80.8	**62.2**	95.2	47.5	**73.5**

Model	Natural Language Inference							Sentence Completion				Coref. Resol.			WSD	Total Avg.
	RTE	**CB**	**AN. R1**	**AN. R2**	**AN. R3**	**EsNLI**	**Avg.**	**COPA**	**Hellasw.**	**StoryC.**	**Avg.**	**Winogr.**	**WSC**	**Avg.**	**WiC**	
mT5	47.3	50.6	32.8	33.4	33	33.9	38.5	56	26	49.9	44	49.2	63.5	56.4	50	40.5
mT-En	81.3	85.1	40.5	36.4	40.2	34.4	53	85.3	34.9	94.7	71.6	61.6	67	64.3	52.6	56.9
mT-Bi	83.9	**88.1**	43.2	36.8	41.6	43.6	56.2	85.7	36	94	71.9	61.7	68.9	65.3	50.8	57.8
mT-Bi+	84.4	86.3	42.4	36.5	40.1	63.5	58.9	**86**	32.5	**94.9**	71.1	**65.4**	**69.2**	**67.3**	51.7	58.8
mT-Bi+-CT	**85.4**	87.5	**43.5**	**37.8**	**43.1**	**73.3**	**61.8**	85.7	**38.3**	94.5	**72.8**	63.1	63.8	63.5	**61.3**	**61.6**

Table 13: P3 bilingual instruction tuning performance breakdown. The best comparable performances are **bolded** and second best underlined.

Model	Sentiment				Avg.	Summarization					Avg.	Multiple-Choice QA					Avg.
	Emot.	**Rot. Tom.**	**Ama.**	**IMDB**		**Mul.News**	**Sam.**	**CNN.**	**XSum**	**Giga.**		**Dream**	**Mul.RC**	**PiQA**	**QASC**	**RACE**	
mT-Ko	35.8	67.7	84.9	75.7	66	**7.8**	11	**15.1**	**11.9**	**22.9**	**13.7**	59.1	65.3	57.2	74.6	44.2	60.1
mT-Ko-Trans	46.4	70.7	**91.2**	88.1	74.1	3.7	20	13.6	9.9	19.1	13.3	59.1	**79.8**	55.8	85.7	48.8	65.8
mT-Ko-CI	**49.4**	**77.8**	**91.2**	**89.6**	**77**	6.7	**20.2**	13.1	10.3	18.3	**13.7**	**69**	**79.8**	**58.2**	**91**	**54.6**	**70.5**

Model	Natural Language Inference						Avg.	Sentence Completion			Avg.	Coref. Resol.		Avg.	WSD	Total Avg.
	RTE	**CB**	**AN. R1**	**AN. R2**	**AN. R3**	**EsNLI**		**COPA**	**Hellasw.**	**StoryC.**		**Winogr.**	**WSC**		**WiC**	
mT-Ko	65.3	75	**36.2**	34	**37.5**	37.5	47.6	63.7	28.9	59.1	50.6	**51.9**	43.3	47.6	50.8	48.1
mT-Ko-CI-Trans	**79.5**	78	31.1	**35**	32.5	52	51.4	**70.3**	27.7	64.1	54	50.3	47.5	48.9	52.3	51.4
mT-Ko-CI	**79.5**	**79.2**	33.6	33.8	33.0	**53.3**	**52.2**	63.3	**35.4**	**67.3**	**55.3**	51.6	**61.9**	**56.8**	**53.3**	**54.1**

Table 14: P3 zero-shot cross-lingual generalization performance breakdown. The best comparable performances are **bolded** and second best underlined.

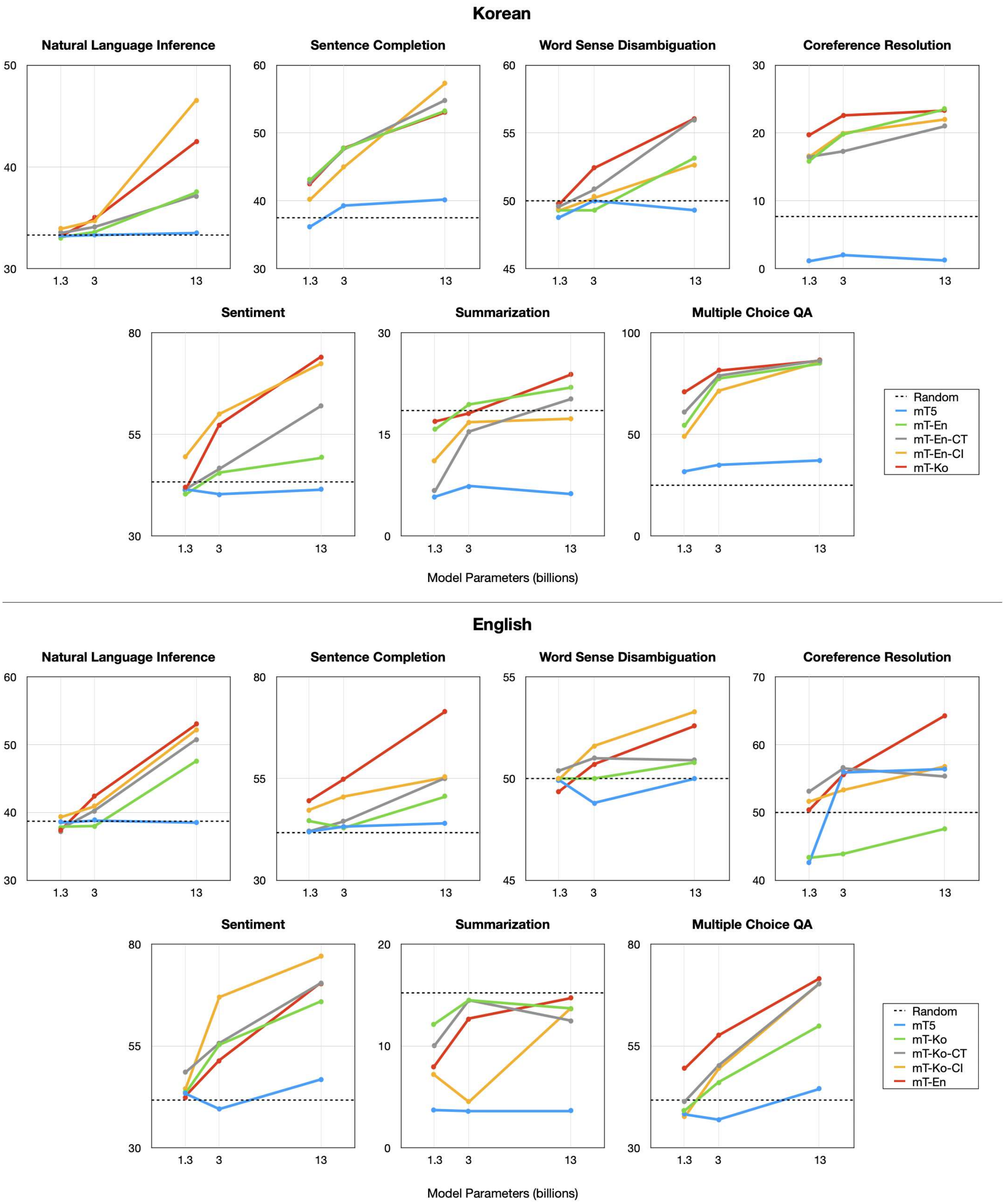

Figure 7: Model performance vs. size. performance breakdown. The random line represents the average score random choice in the options list for classification tasks, and the ROUGE-L score of a copy of input for generation tasks.