How Pre-training Shapes Cross-Lingual Transfer from Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Multilingual language models are pre-trained on uneven amounts of data across languages and can therefore enter reinforcement learning (RL) with different levels of language-specific exposure. RL can further improve their reasoning beyond its training language, but the magnitude of these gains varies across languages. How does target-language pre-training exposure shape cross-lingual transfer from RL? We study this question in a controlled English–Korean arithmetic reasoning setting by varying the amount and difficulty coverage of Korean pre-training data, then applying RL in English, Korean, or both. Cross-lingual transfer is more sensitive to target-language difficulty coverage than to additional data within an already covered range. At similar Korean corpus sizes, broader difficulty coverage yields larger Korean gains from English RL, whereas increasing the Korean corpus ratio with the fixed coverage yields limited additional improvement. Mid-training on previously unseen Korean difficulties improves Korean performance after English-only RL. With a 1% additional budget of mid-training, pass@128 within the RL difficulty range approaches that of a reference pre-trained from scratch with the same difficulty range in both languages. Finally, given the resulting target-language coverage, RL language and difficulty shape where later reasoning gains generalize. Direct Korean RL generalizes more strongly to harder problems beyond the RL range than English-only transfer, while different Korean RL ranges favor different evaluation difficulties. These results show that target-language coverage established through pre-training and mid-training shapes how effectively later reasoning gains transfer across languages, beyond what is reflected in accuracy at the start of RL alone.

1 Introduction

Multilingual pre-training corpora contain substantially different amounts of data across languages (Xue et al., 2021; Ettinger et al., 2025; Abd et al., 2026). Models trained on these corpora can therefore enter subsequent reinforcement learning (RL) with extensive exposure to some languages but limited exposure to others. RL can further improve their reasoning performance (Shao et al., 2024), and these improvements can extend beyond the language used for RL. However, the magnitude of these gains varies across evaluation languages and training configurations (Huang et al., 2026). Understanding this variation calls for examining not only how RL is performed, but also what the model encountered in each language during pre-training. We therefore ask: *How does pre-training shape subsequent RL gains in a multilingual setting with uneven language exposure?*

Existing studies examine how cross-lingual reasoning gains vary with the pre-trained model and post-training configuration (Huang et al., 2026; Dobler et al., 2026; Dhaliwal et al., 2026). However, comparisons among existing pre-trained models make it difficult to isolate how language-specific pre-training exposure contributes to these differences. A complementary line of controlled reasoning research provides a way to study this dependence more directly. Zhang et al. (2026) show that generalization after RL depends jointly on prior pre-training exposure and subsequent training difficulty, and that an intervening mid-training stage can modify these dynamics. Their setting, however, does not address uneven exposure across languages or the resulting cross-lingual transfer. We extend this perspective to a multilingual setting by considering target-language data availability and difficulty coverage. At approximately matched target-language corpus sizes, we vary the range of reasoning difficulties covered during pre-training. When both languages share the same pre-training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

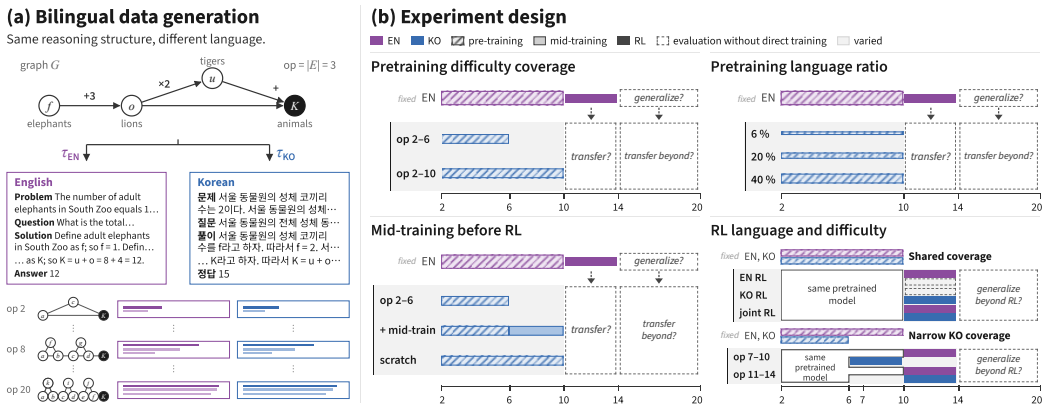


Figure 1: Experiment design overview. (a) English and Korean problems are generated from shared reasoning structures with different language templates. (b) Four experiments examine Korean pre-training difficulty coverage and corpus ratio, mid-training, and RL language and difficulty.

difficulty range, we vary the amount of available target-language data. We investigate how these conditions shape both direct RL gains and cross-lingual transfer. Beyond studying the original pre-training conditions, we also ask whether additional training before RL can mitigate the effects of limited target-language coverage.

We study these dependencies in a controlled English–Korean arithmetic reasoning setting, where shared reasoning structures are expressed in either language and operation count provides a controllable measure of difficulty (Zhou et al., 2025). English receives broad pre-training coverage, while Korean exposure varies in corpus size and covered difficulty range. Holding English-only RL fixed, we first examine how these pre-training conditions affect subsequent cross-lingual transfer. We next test whether missing Korean difficulty coverage can be introduced through mid-training before RL. Given the resulting pre-RL coverage, we then vary RL language and difficulty to examine how later gains generalize within and beyond the RL range.

Contributions. Our study provides a controlled analysis of how language-specific pre-training exposure shapes subsequent cross-lingual reasoning gains from RL. We find that cross-lingual transfer depends more strongly on target-language difficulty coverage than on additional data within an already covered range. At approximately matched Korean corpus size, broader difficulty coverage yields larger Korean gains from English-only RL, whereas increasing the Korean corpus ratio with coverage fixed provides limited additional improvement. We further show that this coverage need not be established entirely during initial pre-training. With a small additional mid-training budget, existing models achieve Korean pass@128 after English-only RL close to the shared-coverage reference within the RL difficulty range. Given the resulting pre-RL coverage, RL language and difficulty determine where later gains generalize. Direct target-language RL provides stronger generalization beyond the RL range than cross-lingual transfer alone, while the effect of RL difficulty depends on its position relative to prior target-language coverage.

2 Experimental Methodology

As illustrated in Figure 1(a), we use generated English and Korean arithmetic problems to study how language-specific pre-training exposure affects reasoning after reinforcement learning (RL). Operation counts define the difficulty ranges used in training and evaluation.

2.1 Dataset

For data generation, we generate arithmetic word problems using GSM-Infinite (Zhou et al., 2025), following Zhang et al. (2026). Dependency graphs specify the arithmetic relations, which are rendered in English or Korean using language-specific templates and entity names. Human reviewers checked the templates for linguistic naturalness. Each example contains a problem, a question, a

step-by-step solution, and a final answer, with the problem and solution in the same language. Complexity is indexed by the generator's operation count, denoted by op. We remove duplicate Korean examples within each op group using the problem, question, and solution text.

For the bilingual setting, the mC4 pre-training corpus, built from Common Crawl, contains more English than Korean text (Xue et al., 2021). We model this imbalance in an English–Korean setting to test whether learning in a data-rich language benefits reasoning in a less-exposed language. Using two languages allows us to vary exposure and difficulty within the same family of arithmetic problems. English receives broad pre-training coverage, while Korean exposure varies in amount and operation-count range. In the English-dominant comparisons, English accounts for a larger share of the available pre-training corpus tokens and Korean is the less-exposed language. For cross-lingual RL comparisons, source and target denote the RL training and transfer evaluation languages, respectively.

2.2 Task setup

We examine whether reasoning learned in one language improves performance in the other and how this transfer depends on training exposure. The comparisons vary language-specific pre-training amounts and operation-count ranges, as well as the language and difficulty range used for RL. English and Korean RL are evaluated in both languages to compare direct training with cross-lingual transfer. Following Zhang et al. (2026), evaluation covers operation counts within and beyond the ranges used in pre-training and RL.

We use pass@128 to assess the range of reasoning problems a model can solve within 128 sampled solutions, following prior analyses of RL generalization (Chen et al., 2026; Zhang et al., 2026). We also report pass@1 to measure single-sample accuracy. Scores are based on final-answer correctness and computed using the pass@k estimator of Chen et al. (2021). Results are reported by language and operation count. Band averages give equal weight to all operation counts within ID (op2–10), Mid-OOD (op11–14), and Hard-OOD (op15–20). In the shared-coverage setting, these bands correspond to the pre-training range, the RL range, and difficulties beyond both. Evaluation details are provided in Appendix A.

2.3 Training setup

Our main experiments use 100M-parameter, decoder-only Transformer models trained from scratch. The bilingual models share an English–Korean tokenizer. We measure training tokens with this tokenizer, since token lengths depend on both language and tokenizer (Petrov et al., 2023). Architecture details and optimization settings are provided in Appendix A.

Pre-training minimizes next-token prediction loss on the generated problems and solutions (Kaplan et al., 2020). Following the data-rich setting of Zhang et al. (2026), the main comparisons use a nominal pre-training budget of approximately 112 tokens per parameter, above the Chinchilla reference of about 20 (Hoffmann et al., 2022). We vary the number of distinct training examples, language composition, and operation-count distribution. Dataset size is distinguished from training exposure, which includes repeated presentations (Muennighoff et al., 2023).

Mid-training continues next-token prediction between pre-training and RL (Zhang et al., 2026; Runwal et al., 2026). In Section 4, we use this stage to expose the model to target-language difficulty levels missing from its pre-training data, while replaying data from both languages. We vary the additional token budget and compare performance after RL. Data composition and optimization settings are provided in Appendix C.

We initialize RL from the pre-trained or mid-trained checkpoint and use Group Relative Policy Optimization (GRPO) (Shao et al., 2024). The reward is 1 for a correct final answer and 0 otherwise. The main comparisons use groups of eight sampled solutions per prompt. We run English-only, Korean-only, or joint English and Korean RL over specified operation ranges and evaluate each model in both languages. Additional training details are provided in Appendix A.

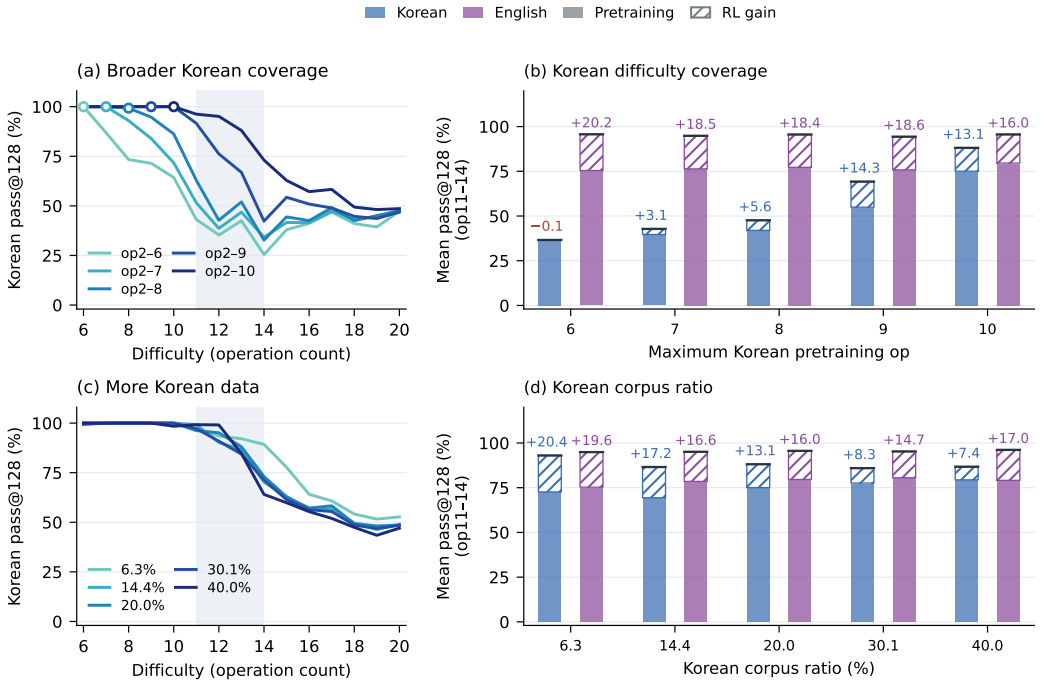


Figure 2: Cross-lingual transfer from English-only RL. Panels (a,b) vary Korean pre-training difficulty coverage, and (c,d) vary the Korean corpus ratio. In (a,c), curves show Korean post-RL pass@128, and shading marks the RL range (op11–14). Open circles in (a) mark Korean pre-training boundaries. In (b,d), pass@128 is averaged equally over op11–14. Solid bars show pre-training scores, dark caps show post-RL scores, and hatching with signed labels shows changes in percentage points. Results are means over three seeds.

3 Pre-Training Conditions Shape Cross-Lingual Transfer

Multilingual models can enter RL with uneven language-specific coverage established during pre-training. Prior work shows that language-specific data availability and allocation affect multilingual and cross-lingual performance (He et al., 2025; Blevins & Zettlemoyer, 2022; Longpre et al., 2025). However, it remains unclear which properties of target-language pre-training are most relevant to the gains transferred from subsequent RL in another language. We focus on two aspects of this exposure, the range of reasoning difficulties represented in the data and the amount of available data. By varying these factors in separate Korean pre-training comparisons while keeping subsequent English-only RL fixed, we examine whether cross-lingual transfer is more closely associated with data amount or with target-language difficulty coverage.

Experimental design. All models use the same English pre-training corpus covering op2–10 and the same pre-training update budget, followed by English-only GRPO on the same op11–14 dataset with a fixed RL budget. We select this slightly harder RL range following Zhang et al. (2026), who report the strongest generalization when applying RL to problems slightly harder than those seen during pre-training. The coverage comparison keeps the available Korean corpus size approximately fixed while extending the Korean pre-training range from op2–6 to op2–10. The data-amount comparison instead fixes the Korean range at op2–10 and increases the available Korean corpus size while holding the English corpus fixed. The comparisons share the op2–10 checkpoint at a 20.0% Korean corpus ratio.

For the data-amount comparison, Korean corpus ratio denotes the fraction of available English–Korean corpus tokens in Korean, counted once, and ranges from 6.3% to 40.0%. Because RL contains only English problems, the Korean change from each pre-trained checkpoint measures cross-lingual transfer, while the English change measures the source-language gain. Other model

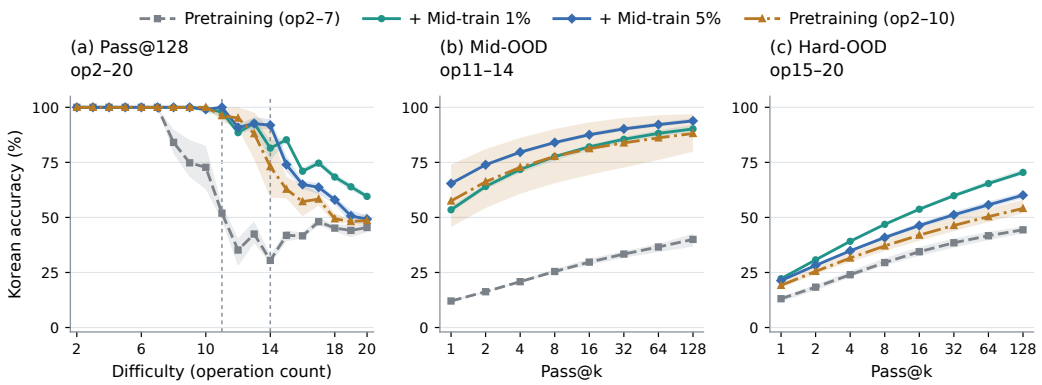


Figure 3: Korean performance after English-only RL. Mid-training extends Korean coverage from op2–7 to op2–10 with additional budgets of 1% and 5%. (a) Pass@128 by operation count. Vertical dashed lines at op11 and op14 mark the RL training range. (b,c) Pass@k averaged equally over all operation counts in Mid-OOD (op11–14) and Hard-OOD (op15–20). Mid-training budgets are percentages of the original nominal pre-training budget. The shared-coverage reference uses the original pre-training budget from scratch. Lines show means over three runs, and ribbons in the corresponding colors span the minimum and maximum scores.

and optimization settings follow Section 2, and Appendix B provides corpus and evaluation details together with the full per-operation results.

Results. At approximately matched Korean corpus size, broader Korean difficulty coverage is associated with substantially larger cross-lingual gains from English RL. Korean gains over op11–14 remain small when pre-training ends at op6–8 and become larger when coverage extends to op9–10, while the corresponding English gains remain similar across conditions (Figure 2(a,b)). With shared op2–10 coverage, the mean Korean and English gains over op11–14 are comparable at 13.1 and 16.0 percentage points, respectively, even though these difficulties were absent from Korean pre-training and RL uses only English problems. This comparison suggests that the range of difficulties represented in the target language can affect subsequent transfer even when the available target-language corpus size is similar. We repeat the coverage comparison at three model scales, labeled 100M, 300M, and 500M (Appendix D). Across all three scales, English-only RL yields substantial Korean pass@128 gains on op11–14 with shared English–Korean op2–10 coverage, while gains remain small with unbalanced coverage (Korean op2–6).

A different pattern appears when English and Korean share the same pre-training difficulty range, op2–10. Increasing the Korean corpus ratio from 6.3% to 40.0% does not consistently improve Korean pass@128 after English-only RL (Figure 2(c,d)).

The contrast between the two comparisons suggests that cross-lingual transfer is more sensitive to whether the target language is represented at relevant reasoning difficulties than to increasing data density within an already covered range. This pattern is consistent with the hypothesis that sparse Korean examples at relevant difficulties provide target-language support for improvements learned in English to transfer to Korean.

Transfer is more limited beyond the English RL range. At op16, English pass@128 gains exceed Korean gains across both comparisons (Tables B1–B4). For the shared-coverage condition with a 20.0% Korean corpus ratio, the mean gains are 55.2 percentage points in English and 6.4 points in Korean.

The main finding is:

> *Target-language difficulty coverage established during pre-training provides a bridge for later reasoning gains to transfer across languages.*

4 MID-TRAINING CAN EXTEND TARGET-LANGUAGE COVERAGE

Section 3 shows that target-language difficulty coverage established before RL can provide a bridge for subsequent cross-lingual transfer. Whether this coverage must be acquired during initial pre-training is less clear. Prior work shows that an intervening mid-training stage can modify subsequent reasoning generalization (Zhang et al., 2026; Runwal et al., 2026). We therefore introduce Korean difficulties missing from pre-training through mid-training, while keeping subsequent English-only RL fixed, and test whether this later exposure allows an existing model to approach the post-RL performance of a model pre-trained with shared coverage.

Experimental design. All conditions use the same English-only GRPO setup and budget on op11–14 as in Section 3. We vary how Korean coverage through op10 is established before RL. The mid-training conditions start from a model pre-trained on English op2–10 and Korean op2–7. Mid-training extends the Korean range by continuing next-token prediction on Korean op8–10, mixed with replayed Korean op2–7 and English op2–10 data. We vary the additional mid-training budget from 1% to 20% of the original nominal pre-training budget. Korean coverage therefore extends to op10 before RL, while op11–14 remains unseen in Korean.

The unbalanced-coverage reference applies English-only RL directly to the same starting model without mid-training. The shared-coverage reference is pre-trained from scratch on English and Korean op2–10, with the Korean op8–10 difficulties included from the start. This reference uses the original nominal pre-training budget. Figure 3 reports Korean performance after RL across evaluation sample budgets, using the Mid-OOD and Hard-OOD bands defined in Section 2.2. Appendix C gives the replay mixture, optimization settings, run counts, and full results after RL.

Results. Introducing the missing Korean difficulties through mid-training improves Korean performance after English-only RL. With a 1% additional mid-training budget, Korean Mid-OOD pass@128 reaches 90.2%, compared with 40.0% without mid-training and 88.1% for the shared-coverage reference trained from scratch (Figure 3(b)). The 1% condition also remains close to the shared-coverage reference across the evaluated sample budgets on Mid-OOD. On Hard-OOD, beyond the English RL range, it achieves higher mean post-RL pass@128 than both references (Figure 3(c)). Mid-training also improves Korean pass@128 on op11–14 after English-only RL at all three tested model scales. In this comparison, mid-training adds the missing Korean op7–10 difficulties to models pre-trained with unbalanced coverage (Korean op2–6), using approximately 1% of the original pre-training budget (Appendix D).

Across the tested mid-training budgets, post-RL Mid-OOD pass@128 remains similar, while pass@1 increases with additional mid-training (Figure 3(b); Table C1). On Hard-OOD, post-RL pass@128 decreases as the mid-training budget increases (Figure 3(c); Table C1).

The comparison in Section 3 concerns coverage acquired during pre-training. Here, supplying missing Korean difficulties after pre-training also improves performance after English-only RL. This dependence on exposure before RL is consistent with the mid-training findings of Zhang et al. (2026) and the effects of data composition reported by Runwal et al. (2026).

The main finding is:

> *Mid-training can add missing target-language difficulty coverage after pre-training and establish a bridge for subsequent cross-lingual transfer.*

5 RL LANGUAGE AND DIFFICULTY SHAPE GENERALIZATION FROM PRE-RL COVERAGE

Section 3 establishes that target-language pre-training coverage affects cross-lingual RL gains, and Section 4 shows that extending this coverage through mid-training improves Korean performance after English-only RL. The limited gains beyond the English RL range in Section 3 motivate examining how RL language and difficulty interact with pre-RL coverage. We first compare RL languages under shared coverage, then vary Korean RL difficulty under unbalanced coverage. We further test whether extending the missing coverage through mid-training changes performance under the same joint RL configuration.

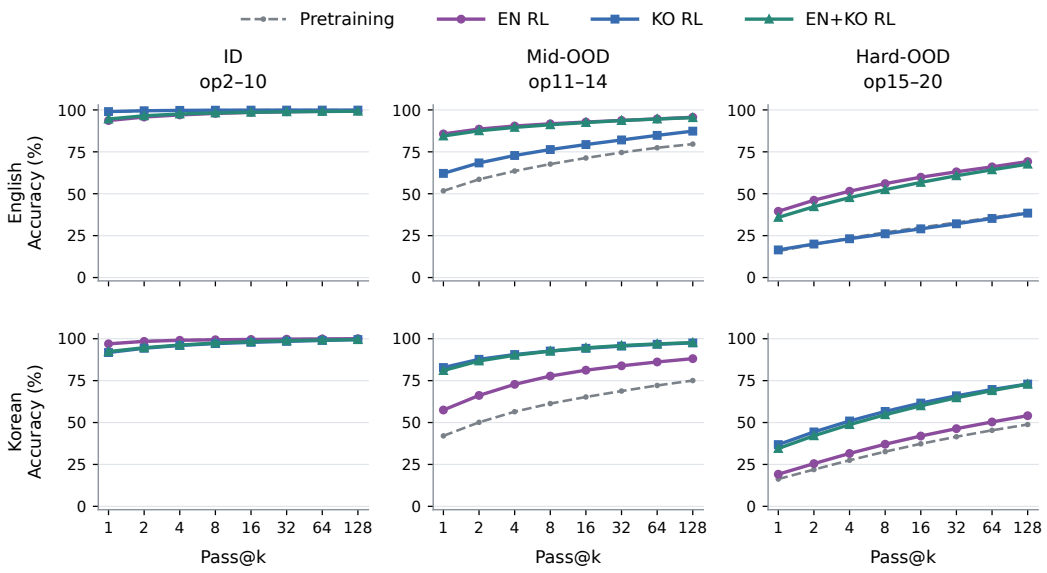


Figure 4: Pass@k after English-only, Korean-only, and 50:50 joint RL from the same three pre-trained models. All RL uses op11–14 and 400k prompt presentations, split equally between languages for joint RL. Scores average all operation counts equally within ID (op2–10), Mid-OOD (op11–14), and Hard-OOD (op15–20). Gray dashed curves show pre-training scores. Curves show means over three runs. Per-operation scores are in Tables E2 and E3.

Experimental design. All conditions use the same total number of training prompt presentations and follow the training configuration in Section 2. Single-language RL allocates the full prompt budget to one language, while joint RL splits it equally between English and Korean.

For the RL-language comparison, we start from the three English-dominant models in Section 3, with a 20.0% Korean corpus ratio and op2–10 coverage in both languages. English-only, Korean-only, and joint RL start independently from each model and use op11–14. Figure 4 compares mean pass@k over three runs before and after RL using the evaluation bands defined in Section 2.2.

To examine RL difficulty under uneven language-specific coverage, we use the three models from Section 3 pre-trained on English op2–10 and Korean op2–6. Each model initializes all RL conditions. Korean-only RL uses either the nearby op7–10 range, immediately beyond Korean pre-training coverage, or the farther op11–14 range. Joint RL fixes English at op11–14 and varies only the Korean range between op7–10 and op11–14. Table 1 reports three-run means, evaluating Korean on both candidate ranges and on op15–20, which lies beyond every tested RL range. English is evaluated on op11–14 and op15–20.

Following Section 4, we introduce Korean op7–10 to each op2–6 model using a 1% additional mid-training budget. We then apply the same joint English–Korean RL on op11–14 to the original models, their mid-trained counterparts, and three separately pre-trained shared-coverage op2–10 references. Appendix E.2 gives the mid-training mixture and full results. Appendix E.3 reports a supplementary budget sweep before joint RL starting from Korean op2–7 pre-training.

Results. Under shared difficulty coverage, Korean-only RL yields higher Korean Mid-OOD pass@1 and pass@128 than English-only RL (Figure 4). The difference is larger at pass@1 (82.8% versus 57.5%) than at pass@128 (97.5% versus 88.1%). This pattern is consistent with a contribution from improved sampling efficiency under direct Korean RL (Chen et al., 2026). Beyond the RL range, Korean-only RL reaches 73.0% pass@128 on Hard-OOD, compared with 54.1% after English-only RL. English-only RL likewise yields higher English Hard-OOD pass@128 than Korean-only RL. Under the same total RL prompt budget, joint RL approaches English-only RL on English and Korean-only RL on Korean at pass@128.

Table 1: RL difficulty under unbalanced pre-training. Scores are mean pass@1/pass@128 (%) over three runs, rounded to whole percentages. Operation counts receive equal weight within each column. English pre-training covers op2–10. The Korean op2–6 conditions share the same three pre-trained models; the final row uses separately pre-trained shared-coverage models with comparable Korean corpus size and the same original pre-training budget. Mid-train gives the additional budget. A dash denotes no training at that stage or in that RL language.

		RL range		Korean (@1/@128)			English (@1/@128)	
Pre-training	Mid-train	KO	EN	op7–10	op11–14	op15–20	op11–14	op15–20
KO op2–6	-	-	-	32 / 68	11 / 37	12 / 42	47 / 75	15 / 40
KO op2–6	-	7–10	-	81 / 91	25 / 58	16 / 44	53 / 77	15 / 39
KO op2–6	-	7–10	11–14	78 / 92	25 / 59	16 / 46	82 / 95	28 / 64
KO op2–6	-	11–14	-	45 / 79	26 / 64	20 / 46	54 / 79	16 / 42
KO op2–6	-	11–14	11–14	45 / 80	24 / 64	19 / 47	82 / 95	29 / 62
KO op2–6	1%	11–14	11–14	77 / 96	54 / 87	25 / 70	81 / 96	29 / 64
KO op2–10	-	11–14	11–14	93 / 100	81 / 98	35 / 73	84 / 95	36 / 68

When Korean pre-training covers only op2–6, assigning the nearby op7–10 range to Korean RL yields much higher accuracy on these previously unseen Korean difficulties than assigning op11–14 (Table 1). This advantage appears in both Korean-only and joint RL. For joint RL, Korean Mid-OOD pass@1 and Hard-OOD pass@128 remain similar across the two placements, as does English performance. Farther Korean RL yields higher Korean Mid-OOD pass@128.

Both joint placements remain below the shared-coverage reference on Mid-OOD and Hard-OOD. Introducing the Korean difficulties missing from pre-training (op7–10) through 1% mid-training before the same joint op11–14 RL raises mean Korean pass@128 from 64.3% to 86.9% on Mid-OOD and from 47.1% to 69.5% on Hard-OOD. The resulting Hard-OOD score is 3.4 percentage points below the shared-coverage reference, while the Mid-OOD gap is 10.8 points. Korean pass@1 also improves, with larger gaps to the reference on both bands.

These results support selecting RL difficulty relative to the target language's pre-training coverage. The same operation range can be covered in one language and unseen in another. In our unbalanced-coverage setting, op7–10 is already covered in English but lies immediately beyond Korean pre-training, so the two languages enter RL with different exposure to these difficulties. This language-specific perspective is consistent with prior findings that RL generalization depends on the relation between training difficulty and prior exposure (Zhang et al., 2026).

The main finding is:

> *Given the coverage established before RL, RL language and difficulty shape where later reasoning gains generalize.*

6 Related Work

The role of pre-training in RL generalization. Reinforcement learning (RL) has improved language-model performance across reasoning tasks and domains (Guo et al., 2025; Huan et al., 2026). Chen et al. (2026) attribute the gains they observe to improved sampling efficiency, as RL makes correct solutions already available to the pre-trained model more likely to be sampled. They report higher pass@1 with comparable or lower pass@k at large sampling budgets. Wu et al. (2025) also report higher pass@1 alongside fewer problems solved under larger sampling budgets. ProRL improves pass@k and solves some problems that the pre-trained model did not solve in the reported evaluation (Liu et al., 2026). These differing findings raise the question of when RL improves the use of knowledge acquired during pre-training and when it extends the model's reasoning ability.

Isolating the role of pre-training is difficult with existing models because their pre-training composition is not independently controlled. Synthetic reasoning tasks allow controlled comparisons of prior training exposure (Ye et al., 2025). In a string-transformation setting, Yuan et al. (2026) show that RL learns new compositions of previously acquired skills and generalizes to more complex compositions. On synthetic programming tasks, Sun et al. (2026) show that RL can learn solutions absent from responses sampled from the pre-trained model when training begins with dense rewards. Using English-language arithmetic tasks and models pre-trained from scratch, Zhang et al. (2026) find pass@128 gains when pre-training leaves room for improvement and RL difficulty is near the limit of what the pre-trained model can solve. Shen et al. (2026) find that lower pre-training validation loss is associated with higher post-RL pass@1 at fixed RL compute in a controlled chess environment. We extend these controlled studies to a bilingual setting to examine how language-specific pre-training exposure shapes the transfer of reasoning gains from RL across languages.

Multilingual reasoning and cross-lingual transfer. Large language models show strong reasoning performance beyond English (Shi et al., 2023; Hu et al., 2025), although accuracy varies across languages and models (Kang et al., 2026). Methods for improving non-English reasoning include cross-lingual prompting (Qin et al., 2023) and model adaptation (Chai et al., 2025; Yoon et al., 2024). Recent approaches use RL, either directly or in combination with supervised fine-tuning (SFT) (Fan et al., 2026; Gupta et al., 2026).

Training in one language can improve reasoning in others, and Huang et al. (2026) find stronger cross-lingual generalization from RL than from SFT in their comparisons. Training on more languages can improve generalization in SFT (Dhaliwal et al., 2026) and parallel RL training (Yang et al., 2025). Park et al. (2025) report GRPO accuracy gains that vary across pre-trained models and target languages. Dobler et al. (2026) observe substantial cross-lingual gains alongside model- and language-dependent regressions on unseen tasks. These comparisons use existing multilingual pre-trained models, so the effects of their language-specific pre-training exposure are not isolated.

Studies of multilingual pre-training evaluate cross-lingual performance after supervised fine-tuning on downstream tasks (Xue et al., 2021) and relate language-modeling performance to data allocation (Chang et al., 2024; He et al., 2025). Longpre et al. (2025) examine cross-lingual transfer during pre-training and subsequent language adaptation in ATLAS. However, how multilingual pre-training conditions affect cross-lingual reasoning gains from subsequent RL remains underexplored. We examine this dependence by varying target-language data availability and difficulty coverage during pre-training, extending coverage through mid-training, and measuring transfer within and beyond the RL difficulty range.

7 Conclusion

Our controlled English–Korean experiments show that pre-training shapes cross-lingual transfer from reinforcement learning by establishing the language-specific difficulty coverage with which a model enters RL. At similar Korean corpus sizes, broader target-language difficulty coverage yields larger Korean gains from English-only RL, whereas adding more Korean data within an already covered difficulty range provides limited additional improvement. Mid-training can extend this coverage after initial pre-training. Introducing the missing Korean difficulties through mid-training improves Korean performance after English-only RL to a level close to the shared-coverage reference within the RL difficulty range. Given the resulting pre-RL coverage, RL language and difficulty shape where later reasoning gains generalize. Direct Korean RL generalizes more strongly to harder Korean problems than English-only transfer, while different Korean RL ranges favor different evaluation difficulties. Together, these results show that cross-lingual transfer from RL is better characterized by which reasoning difficulties are covered in each language before RL and by where RL is applied relative to that coverage than by pre-RL accuracy or target-language data amount alone.

We close with:

> *Pre-training shapes cross-lingual transfer from RL by establishing target-language difficulty coverage. Given that coverage, RL language and difficulty determine where later reasoning gains generalize.*

AI USE STATEMENT

In this work, we used generative AI tools to polish the draft and improve its readability, to create draft of figures, to aid translation, and to help write experiment scripts. All research ideas, the direction of the work, the experimental design, the implementation of the methods, and the interpretation of results were done by the authors without generative AI.

REPRODUCIBILITY STATEMENT

We describe the generation of our English–Korean reasoning datasets in Section 2. Appendix A provides the model architecture, tokenizer, training hyperparameters, and evaluation protocol.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026. doi: 10.48550/ARXIV.2607.02770. URL https://doi.org/10.48550/arXiv.2607.02770.

Terra Blevins and Luke Zettlemoyer. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 3563–3574. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.233. URL https://doi.org/10.18653/v1/2022.emnlp-main.233.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23550–23558, 2025.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Mehak Dhaliwal, Shashwat Chaurasia, Yao Qin, Dezhi Hong, and Thomas Butler. English is not all you need: Systematically exploring the role of multilinguality in llm post-training. *arXiv preprint arXiv:2604.13286*, 2026.

Konstantin Dobler, Federico Scozzafava, Jonathan Janke, Mohamed Ali, and Simon Lehnerer. Grpo beyond english: A large-scale study of grpo in non-english and multilingual settings. *arXiv preprint arXiv:2608.13698*, 2026.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee F. Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, William Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. *CoRR*, abs/2512.13961, 2025. doi: 10.48550/ARXIV.2512.13961. URL https://doi.org/10.48550/arXiv.2512.13961.

Yuchun Fan, Bei Li, Peiguang Li, Yilin Wang, Yongyu Mu, Jian Yang, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, JingBo Zhu, and Tong Xiao. LANG: Reinforcement learning for multilingual reasoning with language-adaptive hint guidance. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43838–43866, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2029. URL https://aclanthology.org/2026.acl-long.2029/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Navya Gupta, Rishitej Reddy Vyalla, Avinash Anand, Chhavi Kirtani, Erik Cambria, Zhengchen Zhang, Zhengkui Wang, Timothy Liu, Aik Beng Ng, Simon See, and Rajiv Ratn Shah. IRIS: Interleaved reinforcement with incremental staged curriculum for cross-lingual mathematical reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22216–22248, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1017. URL https://aclanthology.org/2026.acl-long.1017/.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4257–4273, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Large language models are cross-lingual knowledge-free reasoners. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1525–1542, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.72. URL https://aclanthology.org/2025.naacl-long.72/.

Maggie Ziyu Huan, Yuetai Li, Tianyu Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general LLM capabilities? understanding transferability of LLM reasoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GbOD25IA88.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. In *International Conference on Learning Representations*, volume 2026, pp. 90851–90873, 2026.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Lee. Why do multilingual reasoning gaps emerge in reasoning language models? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31684–31716, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1586. URL https://aclanthology.org/2026.findings-acl.1586/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Advances in Neural Information Processing Systems*, 38:17998–18031, 2026.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Ö. Arik, Chen-Yu Lee, and Sayna Ebrahimi. ATLAS: adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. *CoRR*, abs/2510.22037, 2025. doi: 10.48550/ARXIV.2510.22037. URL https://doi.org/10.48550/arXiv.2510.22037.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. *arXiv preprint arXiv:2506.05850*, 2025.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36: 36963–36990, 2023.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2709, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/.

Bharat Runwal, Ashish Sunil Agrawal, Anurag Roy, and Rameswar Panda. PRISM: Demystifying retention and interaction in mid-training. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=giNBsVVFGt.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training. *arXiv preprint arXiv:2607.16097*, 2026.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hanna Hajishirzi, Nouha Dziri, and Dawn Song. Rl grokking recipe: How does rl unlock and transfer new algorithms in llms? In *International Conference on Learning Representations*, volume 2026, pp. 93190–93220, 2026.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*, pp. 483–498. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Wen Yang, Junhong Wu, Chong Li, Chengqing Zong, and Jiajun Zhang. Parallel scaling law: Unveiling reasoning generalization through a cross-linguistic perspective. *arXiv preprint arXiv:2510.02272*, 2025.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *International Conference on Learning Representations*, volume 2025, pp. 97699–97709, 2025.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. LangBridge: Multilingual reasoning without multilingual supervision. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7502–7522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.405. URL https://aclanthology.org/2024.acl-long.405/.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From f (x) and g (x) to f (g (x)): Llms learn new skills in rl by composing old ones. In *International Conference on Learning Representations*, volume 2026, pp. 147547–147574, 2026.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and RL on reasoning language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TBaUfO9znF.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity?, 2025. URL https://arxiv.org/abs/2502.05252.

A TRAINING AND EVALUATION DETAILS

Table A1: Model.

Parameter	Value
Layers	12
Hidden size	768
FFN size	3,072
Query heads	12
KV heads	2
Activation	SiLU
RMSNorm ϵ	10^{-6}
RoPE base	1,000,000
Tied embeddings	Yes
Vocab. size	1,994
Sequence length	2,048

Table A2: Pre-training.

Parameter	Value
Optimizer	AdamW
Learning rate	10^{-4}
Scheduler	Cosine decay
Min. LR	3×10^{-5}
Warmup	5%
Weight decay	0.1
Clip norm	1.0
Batch size	$32 \times 8 = 256$

Table A3: RL.

Parameter	Value
Learning rate	10^{-6}
Prompt batch	1,024
PPO minibatch	256
Microbatch/device	16
Rollouts/prompt	8
KL weight	10^{-3}
KL estimator	Low-variance
KL in reward	Disabled
Entropy weight	0
Max. prompt	1,024
Max. response	1,024
Temperature	1.0
Top-p	1.0

A.1 MODEL AND TOKENIZER

The main experiments use the same Qwen2.5-style decoder-only architecture with approximately 100M parameters, trained from random initialization. Appendix D gives the larger models used in the comparison across model scales. Table A1 gives the architecture. A shared English–Korean byte-pair-encoding tokenizer has 1,994 tokens, including special tokens. Corpus token counts use this tokenizer, and the maximum sequence length is 2,048.

A.2 PRE-TRAINING AND MID-TRAINING

Pre-training minimizes next-token prediction loss over problems and solutions with AdamW. Table A2 gives the optimization settings, which are also used for mid-training. The comparisons in Section 3 use a nominal pre-training budget of approximately 11.2B training tokens. Available English and Korean training corpus sizes are reported separately in Appendix B. The nominal budget corresponds to 21,353 updates with a global batch of 256 sequences and a maximum sequence length of 2,048. Appendix C gives the mid-training budgets and the budget-matched shared-coverage reference.

A.3 REINFORCEMENT LEARNING

GRPO computes advantages from groups of sampled solutions using binary final-answer rewards based on normalized string matching. A KL loss regularizes the policy against its initialization checkpoint. Table A3 lists the settings used in the main comparisons, with prompt and response limits in tokens. Mid-training followed by English-only and joint RL is described in Appendices C and E.3, respectively.

A.4 EVALUATION

We evaluate 400 problems per operation count in each language. Evaluation samples solutions at temperature 0.7 and top-p 1.0, with a maximum of 1,024 generated tokens. Final answers are compared numerically where possible and by normalized text matching otherwise. We use 128 sampled solutions per problem for the main results and eight for the supplementary pass@1 comparisons in Appendix E.3. Scores average problem-level pass@k within each template and operation count, then weight templates equally. Band averages give operation counts equal weight. Reported run means give runs equal weight; run counts are specified with each comparison.

B PRE-TRAINING DATA AND SUPPLEMENTARY RESULTS

B.1 PRE-TRAINING DATA AND TRAINING CONDITIONS

The coverage comparison in Section 3 varies Korean coverage from op2–6 to op2–10 while keeping the Korean corpus ratio at approximately 20% and the nominal pre-training budget fixed. Each condition combines the same English corpus with approximately 3.0B Korean tokens. Each coverage and data-amount condition is evaluated over three seed runs.

For the data-amount comparison, the English corpus contains approximately 11.8B tokens. The Korean corpora contain approximately 0.8B, 2.0B, 3.0B, 5.1B, and 7.9B tokens, corresponding to Korean corpus ratios of 6.3%, 14.4%, 20.0%, 30.1%, and 40.0% in Figure 2. All checkpoints use the nominal 11.2B-token pre-training budget in Appendix A.

All English-only RL runs use op11–14 with 400,000 prompt presentations, equally allocated across the four operation counts. Other RL settings follow Appendix A. Evaluation follows Appendix A.4, and reported scores are means over the three seeds.

B.2 PER-OPERATION RESULTS

Tables B1–B4 report all evaluated operation counts before and after RL.

Table B1: Korean pass@128 (%) for the coverage comparison in Figure 2. PT and EN RL denote evaluation after pre-training and after English-only RL. Values are means over three seeds. Evaluation sets and aggregation follow Appendix A.4.

KO range	Stage	op2	op3	op4	op5	op6	op7	op8	op9	op10
op2–6	PT	100.0	100.0	100.0	100.0	100.0	81.6	74.1	62.6	53.3
	EN RL	100.0	100.0	100.0	100.0	100.0	86.9	73.3	71.5	64.3
op2–7	PT	100.0	100.0	100.0	100.0	100.0	100.0	89.6	75.6	64.3
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	93.0	83.9	71.7
op2–8	PT	100.0	100.0	100.0	100.0	100.0	100.0	99.3	90.3	70.8
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	99.3	94.7	86.2
op2–9	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
op2–10	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

KO range	Stage	op11	op12	op13	op14	op15	op16	op17	op18	op19	op20
op2–6	PT	43.2	32.8	43.8	27.2	37.3	40.9	45.7	40.4	40.1	47.4
	EN RL	43.2	35.3	42.5	25.4	38.0	41.2	46.9	41.1	39.4	47.3
op2–7	PT	47.8	37.0	45.1	28.7	41.3	41.1	46.3	42.0	43.5	47.1
	EN RL	51.6	38.6	46.8	34.2	41.6	41.6	47.3	43.3	44.4	46.9
op2–8	PT	54.4	38.5	45.4	29.5	46.2	42.7	47.8	44.0	42.6	47.8
	EN RL	62.9	42.7	52.0	32.7	44.4	42.6	49.1	42.6	45.1	47.6
op2–9	PT	74.9	52.1	56.8	36.0	47.4	48.6	49.8	43.7	42.9	48.0
	EN RL	91.6	76.2	66.9	42.3	54.3	50.9	48.9	44.7	43.6	46.9
op2–10	PT	94.4	91.7	67.0	47.1	52.9	50.8	52.4	44.1	44.5	48.3
	EN RL	96.2	95.1	88.0	73.1	62.8	57.2	58.3	49.4	48.1	48.6

Table B2: English pass@128 (%) for the coverage comparison in Figure 2. PT and EN RL denote evaluation after pre-training and after English-only RL. Values are means over three seeds. Evaluation sets and aggregation follow Appendix A.4.

KO range	Stage	op2	op3	op4	op5	op6	op7	op8	op9	op10
op2–6	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0
	EN RL	99.3	98.9	100.0	100.0	100.0	99.5	100.0	99.1	100.0
op2–7	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0
	EN RL	98.0	98.9	99.2	100.0	100.0	98.9	100.0	100.0	100.0
op2–8	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	96.5	99.3	100.0	100.0	100.0	97.9	100.0	100.0	100.0
op2–9	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	99.3	100.0	100.0	99.1	100.0	99.0	100.0	100.0	100.0
op2–10	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	98.4	98.1	100.0	100.0	100.0	98.9	100.0	100.0	99.0

KO range	Stage	op11	op12	op13	op14	op15	op16	op17	op18	op19	op20
op2–6	PT	95.9	84.0	68.1	53.8	33.3	33.2	41.4	37.9	45.6	45.8
	EN RL	98.1	97.7	94.0	92.8	95.5	70.9	60.9	60.6	42.6	49.1
op2–7	PT	98.5	86.7	67.5	52.7	32.4	28.5	39.3	41.1	41.1	44.9
	EN RL	99.1	95.5	93.3	91.6	92.6	70.9	71.3	60.5	44.8	50.0
op2–8	PT	95.6	86.2	70.0	57.0	32.2	32.3	41.8	35.3	39.9	44.2
	EN RL	97.2	97.2	93.3	94.6	95.2	74.7	68.7	65.0	51.5	51.9
op2–9	PT	96.6	86.7	67.2	52.7	29.4	31.7	36.1	35.6	40.0	44.9
	EN RL	97.4	97.2	89.8	93.0	94.4	74.1	66.1	59.3	43.8	50.0
op2–10	PT	96.7	91.2	73.1	57.3	30.6	29.5	37.3	43.6	42.8	49.1
	EN RL	100.0	96.1	94.0	92.3	94.6	84.7	75.2	67.9	42.9	49.8

Table B3: Korean pass@128 (%) for the data-amount comparison in Figure 2. PT and EN RL denote evaluation after pre-training and after English-only RL. Values are means over three seeds. Evaluation sets and aggregation follow Appendix A.4.

Korean corpus ratio (%)	Stage	op2	op3	op4	op5	op6	op7	op8	op9	op10
6.3	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
14.4	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
20.0	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
30.1	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
	EN RL	100.0	100.0	100.0	100.0	99.3	100.0	100.0	100.0	100.0
40.0	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
	EN RL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.4

Korean corpus ratio (%)	Stage	op11	op12	op13	op14	op15	op16	op17	op18	op19	op20
6.3	PT	94.7	76.9	68.4	50.6	54.8	53.1	51.2	47.0	46.0	49.9
	EN RL	97.2	93.5	92.1	89.3	77.7	64.1	60.8	54.2	51.6	52.7
14.4	PT	93.8	80.2	62.6	40.9	49.3	48.2	49.7	42.9	40.2	46.2
	EN RL	99.2	90.2	86.6	70.2	62.9	56.1	56.9	49.5	46.3	49.0
20.0	PT	94.4	91.7	67.0	47.1	52.9	50.8	52.4	44.1	44.5	48.3
	EN RL	96.2	95.1	88.0	73.1	62.8	57.2	58.3	49.4	48.1	48.6
30.1	PT	94.2	88.4	73.7	54.2	55.7	54.3	53.2	46.7	44.6	49.2
	EN RL	97.2	90.9	84.4	71.3	61.4	56.3	55.4	48.6	46.8	48.5
40.0	PT	98.4	97.5	68.9	52.7	54.2	53.4	49.1	46.3	43.0	47.5
	EN RL	99.2	99.1	84.6	64.1	59.9	55.4	52.0	47.4	43.4	47.0

Table B4: English pass@128 (%) for the data-amount comparison in Figure 2. PT and EN RL denote evaluation after pre-training and after English-only RL. Values are means over three seeds. Evaluation sets and aggregation follow Appendix A.4.

Korean corpus ratio (%)	Stage	op2	op3	op4	op5	op6	op7	op8	op9	op10
6.3	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	100.0	100.0	100.0	100.0	100.0	97.6	100.0	100.0	99.4
14.4	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	100.0	98.2	100.0	99.1	100.0	98.9	100.0	99.1	100.0
20.0	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
	EN RL	98.4	98.1	100.0	100.0	100.0	98.9	100.0	100.0	99.0
30.1	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	EN RL	100.0	98.9	99.2	100.0	100.0	98.4	100.0	100.0	100.0
40.0	PT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0
	EN RL	99.3	98.9	98.4	98.5	100.0	98.9	100.0	100.0	100.0

Korean corpus ratio (%)	Stage	op11	op12	op13	op14	op15	op16	op17	op18	op19	op20
6.3	PT	95.9	86.0	65.1	54.2	30.8	31.4	35.2	36.4	42.0	44.9
	EN RL	100.0	95.0	93.1	91.6	95.2	80.7	64.4	56.0	45.4	47.2
14.4	PT	99.1	92.1	70.8	52.2	31.6	28.9	42.6	41.9	45.8	43.5
	EN RL	99.1	94.4	93.0	93.8	93.5	81.7	73.4	70.4	47.9	49.1
20.0	PT	96.7	91.2	73.1	57.3	30.6	29.5	37.3	43.6	42.8	49.1
	EN RL	100.0	96.1	94.0	92.3	94.6	84.7	75.2	67.9	42.9	49.8
30.1	PT	98.2	92.2	76.7	55.2	31.6	34.9	38.4	43.3	41.4	46.8
	EN RL	98.2	96.6	92.5	93.8	93.2	83.4	71.9	63.1	45.0	47.2
40.0	PT	97.4	91.4	72.1	55.6	31.7	33.3	44.8	42.4	41.4	45.4
	EN RL	99.1	99.4	94.0	92.0	90.7	78.2	58.3	57.3	43.1	48.4

C MID-TRAINING DETAILS AND SUPPLEMENTARY RESULTS

C.1 DATA CONSTRUCTION AND TRAINING BUDGETS

Mid-training starts from a model pre-trained with unbalanced coverage (English op2–10 and Korean op2–7) using the nominal 11.2B-token budget. The mid-training budget is split equally among replayed Korean op2–7, new Korean op8–10, and replayed English op2–10 by estimated training tokens. Within each component, operation counts receive equal token budgets. Example counts are selected using measured average token lengths.

Additional mid-training budgets are 0.112B, 0.560B, 1.120B, 1.680B, and 2.240B training tokens, corresponding to 1%, 5%, 10%, 15%, and 20% of the original nominal 11.2B-token training budget. We construct nested training datasets for these budgets and train for one epoch with the optimization settings in Table A2. Evaluation uses the final checkpoint for each budget.

The unbalanced-coverage reference uses the same pre-training configuration without mid-training. The shared-coverage reference in Figure 3 uses the same three English–Korean op2–10 models and English-only RL results as Section 3, with the original nominal 11.2B-token pre-training budget. We also include a *budget-matched shared-coverage reference*, trained from scratch on the original training corpus combined with the largest mid-training dataset. Its nominal 13.4B-token training budget approximately matches the original pre-training plus 20% mid-training budget (1.2×). Table C1 and Figure C1 compare both shared-coverage references.

C.2 ENGLISH RL ACROSS EVALUATION SAMPLE BUDGETS

English-only RL follows the setup and training budget in Appendix B, with hyperparameters in Table A3. Each condition is repeated three times. Mid-training runs start from the same original pre-trained checkpoint. Figure 3 reports performance after English-only RL.

Evaluation follows Appendix A.4.

Table C1 and Figure C1 include all mid-training budgets, extending the 1% and 5% comparison in Figure 3. Table C2 reports post-RL scores at every operation count from op2 to op20.

Adding the missing Korean difficulties through mid-training improves performance on unseen Korean op11–14 problems before RL. With a 1% additional budget, Korean pass@128 reaches 87.9% before RL and 90.2% after English-only RL, compared with 88.1% after RL for the shared-coverage reference.

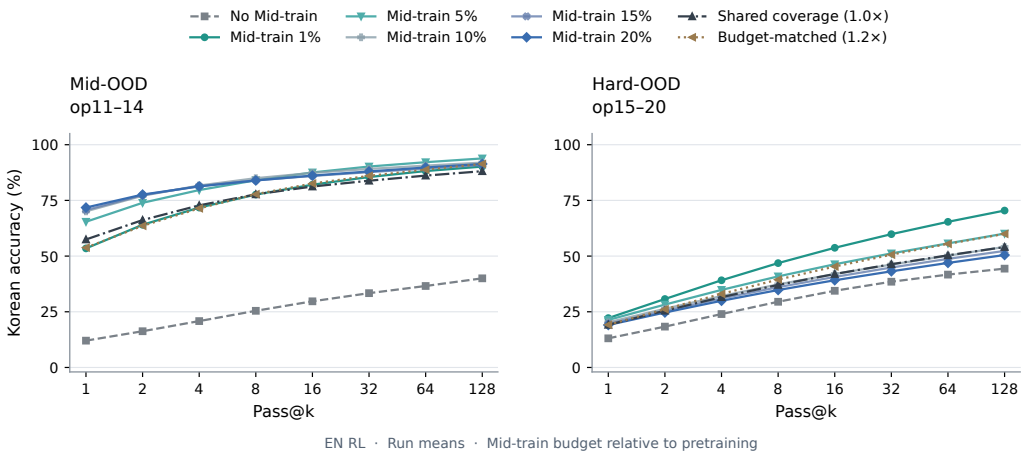


Figure C1: Korean pass@k after English-only RL for all mid-training budgets. Bands average every operation count in Mid-OOD (op11–14) and Hard-OOD (op15–20) equally. Each condition is averaged over three training runs. Shared coverage uses the original training budget (1.0×); the budget-matched reference uses 1.2 times that budget, approximately matching pre-training plus 20% mid-training.

Table C1: Korean pass@k (%) after English-only RL for all mid-training budgets. Mid-OOD and Hard-OOD average every operation count in op11–14 and op15–20, respectively, matching Figure 3. Templates, operation counts, and runs receive equal weights at each aggregation step. Each condition is averaged over three training runs.

Initialization	Band	@1	@2	@4	@8	@16	@32	@64	@128
No Mid-train	Mid-OOD	12.0	16.3	20.8	25.4	29.7	33.4	36.6	40.0
	Hard-OOD	13.1	18.3	24.0	29.5	34.4	38.5	41.7	44.4
Mid-train 1%	Mid-OOD	53.5	64.1	71.8	77.6	82.0	85.5	88.2	90.2
	Hard-OOD	22.2	30.7	39.1	46.8	53.7	59.8	65.4	70.4
Mid-train 5%	Mid-OOD	65.4	73.9	79.6	84.0	87.6	90.2	92.1	93.8
	Hard-OOD	21.3	28.2	34.8	40.9	46.3	51.2	55.7	60.1
Mid-train 10%	Mid-OOD	70.1	77.1	81.7	85.0	87.4	89.1	90.6	92.0
	Hard-OOD	20.3	26.4	32.1	37.3	42.0	46.2	50.2	54.1
Mid-train 15%	Mid-OOD	70.6	77.4	81.4	84.0	86.0	87.7	89.3	90.9
	Hard-OOD	19.5	25.4	31.0	36.0	40.6	44.9	48.7	52.3
Mid-train 20%	Mid-OOD	71.8	77.6	81.3	84.0	86.1	88.0	89.8	91.4
	Hard-OOD	19.1	24.7	29.9	34.8	39.2	43.2	46.9	50.5
Shared coverage (1.0×)	Mid-OOD	57.5	66.2	72.8	77.7	81.2	83.8	86.1	88.1
	Hard-OOD	19.2	25.5	31.6	37.1	42.0	46.4	50.3	54.1
Budget-matched (1.2×)	Mid-OOD	53.8	63.6	71.4	77.8	82.6	86.0	88.8	91.4
	Hard-OOD	19.2	26.2	33.0	39.4	45.3	50.6	55.5	60.0

Table C2: Korean pass@1 and pass@128 (%) after English-only RL at every operation count in Figure 3. Run counts and averaging match Table C1.

	No Mid-train		Mid-train 1%		Mid-train 5%		Shared coverage	
Op	@1	@128	@1	@128	@1	@128	@1	@128
2	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
3	99.1	100.0	94.6	100.0	97.5	100.0	99.0	100.0
4	99.5	100.0	98.8	100.0	99.8	100.0	98.0	100.0
5	97.5	100.0	99.1	100.0	99.9	100.0	99.7	100.0
6	99.1	100.0	96.1	100.0	98.1	100.0	96.8	100.0
7	95.2	100.0	99.7	100.0	99.9	100.0	97.6	100.0
8	56.4	84.1	90.2	100.0	94.3	100.0	93.9	100.0
9	37.6	74.8	96.0	100.0	98.9	100.0	96.2	100.0
10	29.3	72.8	82.2	99.2	90.4	99.1	91.2	100.0
11	16.4	51.9	74.5	98.0	83.1	100.0	79.8	96.2
12	10.8	35.2	61.9	88.5	75.8	90.9	68.7	95.1
13	15.7	42.5	52.3	92.8	60.7	92.6	52.8	88.0
14	5.3	30.5	25.1	81.5	42.0	91.9	28.5	73.1
15	13.1	41.9	34.6	85.3	33.0	74.0	24.1	62.8
16	12.1	41.6	24.0	71.0	24.0	65.0	19.9	57.2
17	17.8	48.1	25.0	74.6	25.0	63.7	23.1	58.3
18	13.0	45.1	19.9	68.3	19.0	58.0	17.1	49.4
19	10.8	44.1	13.6	63.9	12.1	50.7	13.9	48.1
20	11.7	45.4	15.9	59.5	14.9	49.2	16.8	48.6

D MODEL SCALE, COVERAGE, AND MID-TRAINING

We extend the coverage comparison in Section 3 and the mid-training intervention in Section 4 to larger model scales. At each scale, we compare unbalanced pre-training coverage (English op2–10 and Korean op2–6), the same unbalanced coverage followed by mid-training, and shared pre-training coverage (English and Korean op2–10). All conditions then receive English-only RL on op11–14.

D.1 MODELS AND DATA

Table D1 gives the three architectures. We retain the experiment labels 100M, 300M, and 500M; the last model has approximately 562M non-embedding parameters. All models use the same bilingual tokenizer with 1,994 tokens and a maximum sequence length of 2,048. The larger models use maximal update parametrization (μP) during pre-training and mid-training (Yang et al., 2021). Their checkpoints are converted to the standard Qwen2 parameterization before RL and evaluation.

Table D1: Architectures and learning rates for the comparison across model scales. Learning rates apply to both pre-training and mid-training. Other optimization settings follow Table A2.

Setting	100M	300M	500M
Non-embedding parameters	101M	299M	562M
Layers	12	20	20
Hidden size	768	1,024	1,408
FFN size	3,072	4,096	5,632
Query / KV heads	12 / 2	16 / 2	22 / 2
Tied embeddings	Yes	No	No
Learning rate	10^{-4}	4×10^{-4}	4×10^{-4}
Min. LR	3×10^{-5}	1.2×10^{-4}	1.2×10^{-4}

The pre-training corpora are the same as in the corresponding 100M conditions of Appendix B: approximately 11.8B available English tokens from op2–10 and 3.0B Korean tokens from either op2–6 or op2–10. Korean therefore accounts for approximately 20% of available corpus tokens. All model scales use the nominal 11.2B-token pre-training budget defined in Appendix A.2.

Mid-training starts from the unbalanced-coverage models and extends Korean coverage from op2–6 to op2–10 by adding Korean op7–10 examples with replayed Korean op2–6 and English op2–10 examples. The dataset contains 421,639 examples and approximately 0.13B tokens, estimated with the training tokenizer. We use the label Mid-train 1% for this dataset, whose estimated size is approximately 1.1% of the nominal pre-training budget. Each model makes one pass over this dataset using its pre-training learning-rate schedule.

English-only RL uses 200,000 problems, with 50,000 at each operation count from op11 to op14, for two epochs and 400,000 nominal prompt presentations. It uses eight rollouts per prompt and the settings in Table A3. Each model scale and coverage condition has three runs. The 300M and 500M conditions use seeds 51, 52, and 53. At 100M, the unbalanced-coverage and mid-training conditions use seeds 51, 52, and 53, while the shared-coverage reference reuses seeds 42, 2, and 3 from Section 3. Each mid-training run starts from its corresponding unbalanced-coverage pre-trained model.

D.2 EVALUATION AND RESULTS

Evaluation follows Appendix A.4, with band summaries over op11–14. All conditions use the same problem sets through op14, including the reused 100M references from Figure 2.

Under unbalanced coverage, mean post-RL Korean pass@128 on op11–14 is 36.6%, 35.2%, and 35.5% at the 100M, 300M, and 500M scales, respectively. Shared coverage yields 88.1%, 93.5%, and 86.9% at the same scales.

Table D2: Korean op11–14 performance (%) across model scales. Each cell gives the three-run mean and observed minimum–maximum. Unbalanced and shared coverage use Korean op2–6 and op2–10 pre-training, respectively, with English fixed at op2–10. Pre-RL denotes the checkpoint after mid-training for the Mid-train rows and after pre-training otherwise.

Pre-training	Mid-train	KO pass@1		KO pass@128	
		Pre-RL	Post-RL	Pre-RL	Post-RL
100M					
Unbalanced	-	11.1 [8.5, 12.4]	11.3 [9.1, 12.5]	36.7 [31.8, 41.9]	36.6 [29.0, 40.9]
Unbalanced	1%	33.1 [23.4, 38.1]	36.4 [24.0, 45.1]	84.3 [78.7, 88.4]	85.6 [76.8, 93.3]
Shared	-	42.0 [40.5, 45.0]	57.5 [45.6, 74.0]	75.0 [74.0, 76.8]	88.1 [79.8, 97.0]
300M					
Unbalanced	-	13.1 [11.7, 14.1]	13.9 [12.9, 15.8]	33.2 [30.1, 35.4]	35.2 [30.0, 39.8]
Unbalanced	1%	38.4 [30.5, 47.8]	48.6 [39.3, 65.6]	75.0 [65.7, 84.3]	85.6 [80.4, 91.7]
Shared	-	41.7 [34.3, 46.6]	72.1 [69.1, 77.1]	68.4 [60.1, 76.4]	93.5 [92.3, 95.8]
500M					
Unbalanced	-	12.0 [11.2, 13.3]	15.9 [14.6, 17.4]	32.9 [30.4, 35.7]	35.5 [32.8, 40.5]
Unbalanced	1%	37.2 [30.8, 44.2]	44.3 [36.4, 53.6]	80.7 [77.6, 85.3]	86.5 [83.5, 88.3]
Shared	-	38.7 [33.4, 48.6]	63.3 [56.0, 76.5]	64.6 [56.6, 79.0]	86.9 [80.5, 94.9]

Adding the small mid-training dataset raises post-RL Korean pass@128 to 85.6%, 85.6%, and 86.5%, respectively. Table D2 gives the corresponding scores, and Table D3 reports English and Korean pass@1 gains during RL.

Figure D1 shows performance at each operation count.

Table D3: Source-language scores and cross-lingual gains on op11–14 at pass@1. Scores are percentages and gains are percentage points. Each cell gives the three-run mean and observed minimum–maximum.

Condition	EN pass@1		Gain (pp)	
	Pre-RL	Post-RL	ΔKO	ΔEN
100M				
KO op2–6	47.0 [43.9, 49.9]	83.2 [82.6, 83.7]	0.2 [-0.1, 0.6]	36.2 [33.5, 38.8]
KO op2–6 + Mid-train 1%	50.2 [48.6, 52.2]	83.2 [82.8, 83.5]	3.3 [0.6, 7.0]	33.0 [31.0, 34.2]
KO op2–10	51.7 [49.8, 54.4]	85.7 [85.2, 86.0]	15.4 [5.1, 29.0]	34.0 [31.7, 36.1]
300M				
KO op2–6	40.0 [39.3, 40.8]	86.2 [84.2, 88.2]	0.8 [-0.3, 1.7]	46.2 [44.9, 47.5]
KO op2–6 + Mid-train 1%	46.2 [40.6, 54.4]	86.3 [85.7, 87.3]	10.2 [4.2, 17.8]	40.1 [32.9, 45.1]
KO op2–10	44.8 [39.8, 52.6]	88.0 [87.0, 89.3]	30.3 [22.5, 35.7]	43.2 [34.3, 47.8]
500M				
KO op2–6	44.4 [40.1, 50.9]	88.0 [86.1, 90.2]	3.9 [3.3, 4.2]	43.6 [35.2, 50.1]
KO op2–6 + Mid-train 1%	45.3 [40.6, 51.3]	86.9 [85.8, 87.5]	7.1 [5.6, 9.4]	41.7 [34.5, 46.9]
KO op2–10	40.9 [36.6, 48.0]	85.6 [83.8, 87.1]	24.6 [22.7, 27.8]	44.7 [37.9, 50.5]

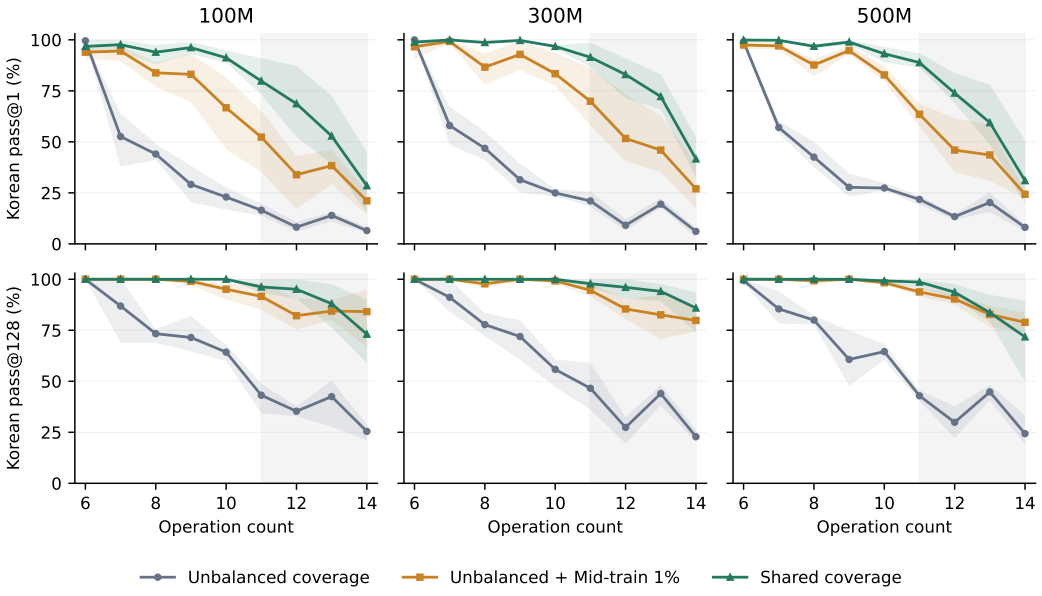


Figure D1: Korean performance after English-only RL across model scales and operation counts. Lines show means over three runs, and shaded ribbons span the observed minimum and maximum. The gray region marks the English RL range, op11–14. Unbalanced and shared coverage use Korean op2–6 and op2–10 pre-training, respectively, with English fixed at op2–10. Mid-training extends Korean coverage from op2–6 to op2–10 before RL.

E RL LANGUAGE AND DIFFICULTY COMPARISONS

This appendix gives the full results for the RL language and difficulty comparisons in Section 5 and a supplementary mid-training budget sweep before joint RL.

E.1 RL LANGUAGE ACROSS EVALUATION SAMPLE BUDGETS

Figure 4 uses the same three pre-trained models and training-token budget as the 20.0% Korean corpus ratio condition with shared op2–10 coverage in Section 3. All RL conditions use a training budget of 400,000 prompt presentations on op11–14. Joint RL divides this budget equally between English and Korean. RL hyperparameters follow Table A3.

Evaluation follows Appendix A.4, using the common problem sets from Figure 2. Reported scores are means over three runs.

Table E1 gives every plotted point. Tables E2 and E3 give pass@1 and pass@128 at every evaluated operation count. English ID pass@1 decreases from 98.9% after pre-training to 93.7% after English RL. Pass@128 remains high at 99.4%, compared with 99.9% before RL.

Table E1: Mean pass@k (%) over three runs for Figure 4. ID, Mid-OOD, and Hard-OOD average every operation count in op2–10, op11–14, and op15–20, respectively, with equal weights. Pre-training denotes the checkpoint before RL. EN RL, KO RL, and EN+KO RL denote English-only, Korean-only, and joint RL on op11–14.

Training	Band	@1	@2	@4	@8	@16	@32	@64	@128
English evaluation									
Pre-training	ID	98.9	99.5	99.7	99.8	99.9	99.9	99.9	99.9
	Mid-OOD	51.7	58.6	63.5	67.7	71.3	74.6	77.4	79.6
	Hard-OOD	15.8	19.8	23.6	26.9	29.9	32.8	35.8	38.8
EN RL	ID	93.7	95.7	97.0	97.9	98.5	98.9	99.2	99.4
	Mid-OOD	85.7	88.6	90.4	91.7	92.8	93.7	94.6	95.6
	Hard-OOD	39.5	46.1	51.5	56.0	59.9	63.1	66.0	69.2
KO RL	ID	99.0	99.5	99.7	99.8	99.9	99.9	99.9	99.9
	Mid-OOD	62.1	68.4	72.8	76.4	79.3	82.1	84.8	87.3
	Hard-OOD	16.5	20.0	23.1	26.1	29.0	32.1	35.2	38.4
EN+KO RL	ID	94.6	96.5	97.6	98.2	98.6	98.9	99.1	99.2
	Mid-OOD	84.4	87.6	89.6	91.1	92.5	93.7	94.6	95.4
	Hard-OOD	35.9	42.3	47.7	52.4	56.8	60.7	64.3	67.6
Korean evaluation									
Pre-training	ID	96.9	98.5	99.1	99.4	99.6	99.7	99.8	99.9
	Mid-OOD	42.0	50.1	56.5	61.3	65.2	68.8	72.1	75.0
	Hard-OOD	16.3	21.9	27.5	32.6	37.3	41.5	45.4	48.8
EN RL	ID	96.9	98.4	99.1	99.4	99.6	99.7	99.9	100.0
	Mid-OOD	57.5	66.2	72.8	77.7	81.2	83.8	86.1	88.1
	Hard-OOD	19.2	25.5	31.6	37.1	42.0	46.4	50.3	54.1
KO RL	ID	91.8	94.3	96.0	97.1	97.8	98.4	99.0	99.5
	Mid-OOD	82.8	87.7	90.6	92.6	94.3	95.6	96.7	97.5
	Hard-OOD	36.8	44.4	50.9	56.6	61.6	65.9	69.7	73.0
EN+KO RL	ID	92.3	94.6	96.2	97.4	98.3	98.8	99.2	99.5
	Mid-OOD	81.0	86.7	90.1	92.6	94.6	95.9	96.9	97.7
	Hard-OOD	34.5	42.1	48.8	54.7	60.0	64.8	69.0	72.9

Table E2: English pass@1 and pass@128 (%) at each operation count for Figure 4. Values are means over three runs. Pre-training denotes the checkpoint before RL. EN RL, KO RL, and EN+KO RL denote English-only, Korean-only, and joint RL on op11–14.

	Pre-training		EN RL		KO RL		EN+KO RL	
Op	@1	@128	@1	@128	@1	@128	@1	@128
2	100.0	100.0	87.6	98.4	100.0	100.0	88.6	97.8
3	99.4	100.0	85.2	98.1	99.3	100.0	89.6	98.9
4	100.0	100.0	93.2	100.0	99.9	100.0	94.9	98.9
5	100.0	100.0	98.3	100.0	100.0	100.0	97.9	100.0
6	99.9	100.0	94.7	100.0	99.9	100.0	95.5	98.9
7	99.6	100.0	94.7	98.9	99.7	100.0	94.4	99.5
8	99.3	100.0	94.9	100.0	99.5	100.0	95.6	100.0
9	98.4	100.0	98.2	100.0	98.3	100.0	97.9	100.0
10	93.9	99.0	96.0	99.0	94.3	99.0	96.8	99.0
11	83.5	96.7	93.8	100.0	88.4	100.0	93.9	99.1
12	59.0	91.2	91.1	96.1	72.8	94.3	90.0	94.4
13	36.6	73.1	81.7	94.0	54.6	85.6	78.5	94.2
14	27.7	57.3	76.1	92.3	32.8	69.5	75.5	93.9
15	17.2	30.6	69.4	94.6	20.9	41.4	65.5	93.9
16	18.0	29.5	61.2	84.7	16.2	32.4	55.2	85.9
17	11.6	37.3	39.5	75.2	14.5	33.2	33.4	71.0
18	13.7	43.6	26.4	67.9	13.3	39.7	22.9	62.9
19	17.8	42.8	23.9	42.9	17.7	39.6	20.7	42.4
20	16.3	49.1	16.8	49.8	16.2	44.4	17.5	49.8

Table E3: Korean pass@1 and pass@128 (%) at each operation count for Figure 4. Values are means over three runs. Pre-training denotes the checkpoint before RL. EN RL, KO RL, and EN+KO RL denote English-only, Korean-only, and joint RL on op11–14.

	Pre-training		EN RL		KO RL		EN+KO RL	
Op	@1	@128	@1	@128	@1	@128	@1	@128
2	100.0	100.0	100.0	100.0	96.9	99.3	97.7	100.0
3	99.4	100.0	99.0	100.0	81.8	100.0	83.9	99.2
4	98.0	100.0	98.0	100.0	89.6	100.0	91.1	100.0
5	99.7	100.0	99.7	100.0	93.7	99.1	94.9	99.1
6	96.9	100.0	96.8	100.0	90.2	98.7	90.6	98.7
7	97.6	100.0	97.6	100.0	94.3	99.3	94.3	100.0
8	94.3	100.0	93.9	100.0	90.6	100.0	90.1	100.0
9	96.7	100.0	96.2	100.0	94.7	100.0	94.3	100.0
10	89.8	99.2	91.2	100.0	94.3	99.1	94.2	98.3
11	75.4	94.4	79.8	96.2	94.3	99.2	92.0	99.2
12	48.0	91.7	68.7	95.1	84.5	100.0	83.3	99.1
13	31.7	67.0	52.8	88.0	80.9	95.9	79.8	93.8
14	13.1	47.1	28.5	73.1	71.5	95.0	69.0	98.7
15	17.3	52.9	24.1	62.8	67.3	92.5	65.8	92.4
16	17.3	50.8	19.9	57.2	39.1	79.9	36.3	79.5
17	20.8	52.4	23.1	58.3	40.1	77.7	37.7	77.5
18	15.2	44.1	17.1	49.4	30.3	67.4	27.0	66.0
19	11.6	44.5	13.9	48.1	22.6	62.0	20.3	64.6
20	15.5	48.3	16.8	48.6	21.6	58.4	20.0	57.5

E.2 RL PLACEMENT UNDER UNBALANCED COVERAGE

The main placement comparison starts from the three unbalanced-coverage models (Korean op2–6) in Appendix B. English covers op2–10, the Korean corpus contains approximately 3.0B available tokens, and pre-training uses the nominal 11.2B-token budget. Each model initializes all RL conditions. The shared-coverage references use the op2–10 pre-training configuration with a 20.0% Korean corpus ratio used in Figure 4. These three models are pre-trained separately with approximately 3.0B available Korean corpus tokens and the same nominal pre-training token budget.

Single-language RL uses 50,000 problems at each of four operation counts for two epochs. Joint RL uses 50,000 problems per language and operation count for one epoch. Every RL condition therefore has 400,000 nominal prompt presentations, with equal allocation by language in joint RL. All runs use eight rollouts per prompt and are evaluated after two epochs for single-language RL or one epoch for joint RL. Other settings follow Table A3. Korean RL uses either op7–10, immediately beyond its pre-training range, or op11–14. English RL uses op11–14.

The mid-training condition uses the Korean op7–10 dataset and training recipe in Appendix D.1, starting from each Korean op2–6 model. Subsequent joint RL uses op11–14 in both languages with the same prompt allocation as the corresponding condition without mid-training.

Evaluation follows Appendix A.4, using the common problem sets from Figure 2. Reported scores are means over three runs.

Tables E4 and E5 give the complete per-operation results for Table 1, with English-only RL from the same Korean op2–6 models included as an additional comparison.

Table E4: Korean pass@1 and pass@128 (%) at every operation count for Table 1 and English-only RL. Values are means over three runs. The additional English-only RL condition starts from the same Korean op2–6 pre-trained models. Joint RL uses English op11–14 and the Korean range in parentheses; mid-training and shared-coverage references use joint op11–14 RL. Evaluation follows Appendix E.2.

Setting	k	2	3	4	5	6	7	8	9	10
Pre-training	1	100.0	98.6	98.8	99.4	99.4	45.0	38.5	22.4	20.9
	128	100.0	100.0	100.0	100.0	100.0	81.6	74.1	62.6	53.3
EN RL (op11–14)	1	100.0	98.5	98.5	99.4	99.5	52.6	44.0	29.1	23.0
	128	100.0	100.0	100.0	100.0	100.0	86.9	73.3	71.5	64.3
KO RL (op7–10)	1	78.7	47.5	58.9	66.7	77.3	97.4	86.2	78.7	60.9
	128	97.7	88.6	89.9	93.8	96.4	99.3	94.4	87.7	84.4
Joint RL (KO op7–10)	1	82.4	47.5	60.1	62.5	70.3	96.0	83.0	74.8	56.8
	128	99.3	86.2	85.5	92.1	95.6	100.0	95.6	89.1	83.5
KO RL (op11–14)	1	78.4	36.4	37.4	39.3	37.1	44.1	41.8	46.7	46.9
	128	94.9	77.2	72.5	83.1	79.2	78.2	78.1	81.3	79.6
Joint RL (KO op11–14)	1	83.6	47.4	31.5	37.6	38.0	42.4	42.4	49.8	45.0
	128	100.0	90.9	61.1	73.1	63.4	74.4	80.7	84.2	80.0
Mid-train 1% + joint RL	1	86.9	58.2	64.6	75.7	65.6	75.5	75.6	82.3	75.2
	128	99.3	88.6	95.4	92.6	91.3	96.5	93.7	100.0	95.1
Shared coverage + joint RL	1	97.7	83.9	91.1	94.9	90.6	94.3	90.1	94.3	94.2
	128	100.0	99.2	100.0	99.1	98.7	100.0	100.0	100.0	98.3

Setting	k	11	12	13	14	15	16	17	18	19	20
Pre-training	1	15.2	9.1	13.6	6.5	12.8	12.0	14.3	11.1	9.4	13.0
	128	43.2	32.8	43.8	27.2	37.3	40.9	45.7	40.4	40.1	47.4
EN RL (op11–14)	1	16.5	8.2	13.9	6.5	12.5	11.8	13.6	11.0	9.4	12.6
	128	43.2	35.3	42.5	25.4	38.0	41.2	46.9	41.1	39.4	47.3
KO RL (op7–10)	1	43.8	22.5	24.3	11.1	16.6	16.1	21.4	15.6	12.5	15.5
	128	74.5	59.4	60.4	36.9	44.9	43.0	50.0	41.2	38.8	45.2
Joint RL (KO op7–10)	1	42.2	20.1	25.1	10.7	16.1	16.0	20.2	14.5	12.9	15.5
	128	76.1	62.2	59.3	36.5	47.3	44.9	52.8	42.9	42.7	48.3
KO RL (op11–14)	1	33.0	23.0	29.4	20.4	20.5	20.4	24.4	19.4	16.0	18.8
	128	67.0	61.9	60.7	64.7	56.7	47.1	50.7	44.0	36.2	42.4
Joint RL (KO op11–14)	1	33.7	17.9	27.5	16.9	19.7	19.2	23.3	18.5	14.8	18.2
	128	67.2	59.5	63.7	66.9	56.4	46.6	50.2	44.2	41.5	43.4
Mid-train 1% + joint RL	1	64.8	57.0	55.3	39.5	42.1	24.1	28.3	21.5	16.4	18.1
	128	90.3	83.9	84.7	88.7	83.7	72.0	73.4	64.6	63.6	59.8
Shared coverage + joint RL	1	92.0	83.3	79.8	69.0	65.8	36.3	37.7	27.0	20.3	20.0
	128	99.2	99.1	93.8	98.7	92.4	79.5	77.5	66.0	64.6	57.5

Table E5: English pass@1 and pass@128 (%) at every operation count for Table 1 and English-only RL. Values are means over three runs. The additional English-only RL condition starts from the same Korean op2–6 pre-trained models. Joint RL uses English op11–14 and the Korean range in parentheses; mid-training and shared-coverage references use joint op11–14 RL. Evaluation follows Appendix E.2.

Setting	k	2	3	4	5	6	7	8	9	10
Pre-training	1	100.0	98.6	100.0	100.0	99.9	99.6	98.4	98.1	92.0
	128	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0
EN RL (op11–14)	1	93.6	93.4	91.7	96.1	97.2	97.3	98.0	95.4	97.0
	128	99.3	98.9	100.0	100.0	100.0	99.5	100.0	99.1	100.0
KO RL (op7–10)	1	100.0	98.5	99.9	100.0	99.9	99.6	98.7	98.1	92.9
	128	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
Joint RL (KO op7–10)	1	93.7	94.1	93.8	98.1	99.1	96.8	98.1	96.6	96.3
	128	99.3	100.0	100.0	100.0	100.0	99.5	100.0	99.1	100.0
KO RL (op11–14)	1	99.9	98.4	99.8	100.0	99.6	99.4	97.9	98.1	91.9
	128	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
Joint RL (KO op11–14)	1	95.7	94.3	94.6	97.9	99.2	96.9	97.6	96.3	96.1
	128	99.3	98.9	100.0	100.0	100.0	99.5	100.0	98.3	99.0
Mid-train 1% + joint RL	1	96.6	95.4	96.2	98.3	99.3	97.3	98.0	95.2	96.2
	128	100.0	100.0	100.0	100.0	100.0	99.5	100.0	99.1	99.0
Shared coverage + joint RL	1	88.6	89.6	94.9	97.9	95.5	94.4	95.6	97.9	96.8
	128	97.8	98.9	98.9	100.0	98.9	99.5	100.0	100.0	99.0

Setting	k	11	12	13	14	15	16	17	18	19	20
Pre-training	1	78.0	48.4	33.5	28.2	13.9	18.1	13.6	15.2	15.3	15.3
	128	95.9	84.0	68.1	53.8	33.3	33.2	41.4	37.9	45.6	45.8
EN RL (op11–14)	1	92.7	88.8	80.4	71.1	54.0	41.0	28.8	18.4	20.0	17.8
	128	98.1	97.7	94.0	92.8	95.5	70.9	60.9	60.6	42.6	49.1
KO RL (op7–10)	1	82.0	60.2	38.8	29.5	14.0	17.1	14.6	15.5	16.5	14.6
	128	95.7	81.4	72.4	57.1	34.1	32.0	38.8	41.9	41.9	45.1
Joint RL (KO op7–10)	1	92.1	88.0	77.3	69.5	50.2	35.7	27.5	18.4	19.2	17.0
	128	98.1	96.6	92.3	93.8	94.7	71.6	58.3	59.4	45.5	52.3
KO RL (op11–14)	1	81.3	59.9	44.2	29.6	13.6	16.4	15.5	15.3	17.9	15.0
	128	95.7	86.6	73.9	59.3	37.7	32.6	44.4	39.2	45.4	49.8
Joint RL (KO op11–14)	1	91.8	86.8	78.7	68.8	50.3	38.0	30.5	17.0	19.9	17.5
	128	97.2	96.6	93.3	93.0	94.4	67.6	60.2	56.8	45.6	48.6
Mid-train 1% + joint RL	1	90.8	85.1	79.0	68.7	49.6	39.4	29.2	19.2	18.7	18.7
	128	99.1	97.1	94.0	91.9	91.1	73.6	61.0	58.9	49.3	49.8
Shared coverage + joint RL	1	93.9	90.0	78.5	75.5	65.5	55.2	33.4	22.9	20.7	17.5
	128	99.1	94.4	94.2	93.9	93.9	85.9	71.0	62.9	42.4	49.8

E.3 MID-TRAINING BUDGETS BEFORE JOINT RL

Table E6 examines how the mid-training budget affects performance after joint RL. Mid-training starts from a model pre-trained with unbalanced coverage (English op2–10 and Korean op2–7) and adds Korean op8–10 using the replay mixture and budgets in Appendix C.1. All conditions then receive 50:50 joint RL on op11–14 with 400,000 prompt presentations, following Appendix E.1. We report pass@1 from individual runs. Evaluation follows Appendix A.4. ID* averages op2, op4, op6, op8, and op10; Mid-OOD averages op11–14; Hard-OOD averages op16, op18, and op20. Tables E8 and E9 give the per-operation scores.

After joint RL, Korean Mid-OOD pass@1 is 53.1% for the unbalanced-coverage reference and 91.0% at 15% mid-training, compared with 88.6% for the shared-coverage reference (Table E6). English Mid-OOD pass@1 remains around 87% across the mid-training budgets, compared with 87.4% for the shared-coverage reference.

On Korean Hard-OOD, joint RL reaches approximately 31–32% pass@1 across the mid-training budgets, compared with 20.8% for the unbalanced-coverage reference and 34.0% for the shared-coverage reference. At 15% mid-training, Korean Mid-OOD exceeds the shared-coverage reference while Hard-OOD remains 2 to 3 points below it. The budget-matched reference reaches 91.7% on Mid-OOD and 38.8% on Hard-OOD (Table E7).

Table E6: Mean pass@1 (%) after 50:50 joint RL on op11–14 with 400k prompt presentations. Mid-training percentages are additional nominal budgets relative to the unbalanced-coverage checkpoint. Both from-scratch references use the original budget; the shared-coverage model is the shared-coverage checkpoint of Section 3 after the same joint RL. Bands follow Appendix E.3; ID* includes difficulties absent from unbalanced Korean pre-training, and Hard-OOD exceeds all training ranges. Results are individual runs.

	English			Korean		
Initialization	ID* op2–10	Mid-OOD op11–14	Hard-OOD op16, 18, 20	ID* op2–10	Mid-OOD op11–14	Hard-OOD op16, 18, 20
Unbalanced pre-training + mid-training						
Mid-train 1%	95.2	86.9	37.5	88.8	80.0	31.8
Mid-train 5%	95.7	86.6	31.1	95.8	87.2	31.4
Mid-train 10%	94.9	86.8	31.2	96.9	88.3	31.8
Mid-train 15%	94.4	87.0	29.0	98.2	91.0	31.0
Mid-train 20%	94.6	87.2	32.1	97.9	90.1	31.1
From-scratch references						
Unbalanced coverage	95.6	86.6	33.8	76.0	53.1	20.8
Shared coverage	94.6	87.4	31.2	95.8	88.6	34.0

Table E7: Mean pass@1 (%) after 50:50 joint RL for the from-scratch references, with the protocol of Table E6. The unbalanced-coverage and shared-coverage references use the original nominal 11.2B-token training budget. The budget-matched reference adds the largest mid-training dataset to the training corpus and uses a nominal 13.4B-token training budget (1.2×). Results are individual runs.

	English			Korean		
Initialization	ID* op2–10	Mid-OOD op11–14	Hard-OOD op16, 18, 20	ID* op2–10	Mid-OOD op11–14	Hard-OOD op16, 18, 20
From-scratch references after joint RL						
Unbalanced coverage	95.6	86.6	33.8	76.0	53.1	20.8
Shared coverage (1.0×)	94.6	87.4	31.2	95.8	88.6	34.0
Budget-matched (1.2×)	96.8	88.8	32.7	96.3	91.7	38.8

Table E8: Per-operation English pass@1 (%) after 50:50 joint RL, corresponding to the band averages in Table E6 and the references in Table E7. Training conditions follow those tables.

	ID*					Mid-OOD				Hard-OOD		
Initialization	2	4	6	8	10	11	12	13	14	16	18	20
Unbalanced pre-training + mid-training												
Mid-train 1%	96.2	96.0	95.7	95.9	92.3	94.4	90.9	82.8	79.7	58.1	35.2	19.2
Mid-train 5%	96.5	97.1	96.1	96.0	92.5	94.5	90.6	82.8	78.7	52.7	24.8	15.9
Mid-train 10%	95.7	95.7	95.5	95.3	92.2	93.9	91.2	83.3	78.8	55.4	22.9	15.3
Mid-train 15%	95.7	94.5	93.8	95.1	92.6	93.9	91.5	83.2	79.5	50.0	21.9	15.0
Mid-train 20%	96.6	94.6	94.5	95.1	92.2	94.5	90.7	83.5	80.0	55.6	25.0	15.8
From-scratch references												
Unbalanced coverage	96.9	97.0	95.8	96.0	92.5	94.0	90.9	82.1	79.3	53.1	30.6	17.7
Shared coverage (1.0×)	94.5	94.3	95.7	95.9	92.8	94.3	91.3	83.9	80.1	54.0	23.6	16.1
Budget-matched (1.2×)	98.4	98.3	97.7	96.9	92.7	96.0	91.6	84.7	83.1	55.9	26.1	16.2

Table E9: Per-operation Korean pass@1 (%) after 50:50 joint RL, corresponding to the band averages in Table E6 and the references in Table E7. Training conditions follow those tables.

	ID*					Mid-OOD				Hard-OOD		
Initialization	2	4	6	8	10	11	12	13	14	16	18	20
Unbalanced pre-training + mid-training												
Mid-train 1%	89.9	82.5	86.1	91.0	94.6	91.2	88.8	73.7	66.4	39.1	34.4	22.0
Mid-train 5%	95.8	94.3	95.2	96.0	97.5	94.5	94.5	82.5	77.3	40.4	33.7	20.2
Mid-train 10%	98.0	96.5	95.8	96.1	97.8	95.9	94.8	84.2	78.3	41.5	34.7	19.2
Mid-train 15%	99.3	98.7	97.2	97.9	98.0	96.5	96.5	87.1	83.8	45.5	30.3	17.2
Mid-train 20%	99.8	98.5	96.4	97.0	97.9	96.6	96.0	85.5	82.2	42.6	32.5	18.3
From-scratch references												
Unbalanced coverage	82.5	73.0	73.2	81.3	70.2	72.3	54.3	42.8	42.9	23.9	21.2	17.1
Shared coverage (1.0×)	95.2	96.1	94.2	95.7	97.7	96.1	94.6	82.9	80.8	44.7	37.0	20.3
Budget-matched (1.2×)	94.8	94.3	97.2	96.8	98.3	97.1	96.3	88.1	85.4	52.9	40.6	22.9