Kowen: Training a Strong Bilingual LLM through Synthetic Data

Anonymous authors
Paper under double-blind review

Abstract

Due to their versatility in a wide coverage of fields and tasks, large language models (LLMs) have surged as proprietary products, often lacking openness in its recipe to replicate the training process. Even though major research movements to replicate and elucidate the English LLMs have been proposed, this does not apply to other languages like Korean. Thus, we present a `Kowen`, an open-source state-of-the-art Korean and English-speaking LLM, trained primarily through the active usage of synthetic data. We first reveal the effect of leveraging multilingual synthetic data from teacher models. Then, we scale the model and data size to train a strong bilingual LLM with the combination of supervised fine-tuning from teacher responses and iterative fine-tuning. We release all details and code for reproducibility.

1 Introduction

Training large language models (LLMs) to diverse languages and cultures is essential in creating a more equitable representation of AI systems. As a drastic portion of text corpus for LLM training constitutes English, considerate research efforts were in need in creating multilingual large language models (MLLMs). Recent advances in scaling data (Ouyang et al., 2022; Dubey et al., 2024; Yang et al., 2024) have enabled the exploration of scaling under-represented multilingual text data as well.

Attempts to develop open post-training recipes with high-quality synthetic datasets closed the gap with the close-recipe models (Tunstall et al., 2024; Bartolome et al., 2024; Lambert et al., 2024; OLMo et al., 2025). Tunstall et al. (2024, Zephyr) proposed the paradigm of applying direct preference optimization (Rafailov et al., 2024, DPO) to enhance instruction-following abilities with synthetic preference dataset, surpassing larger close-recipe models like Llama-2 series (Touvron et al., 2023). Inspired by Zephyr, Bartolome et al. (2024, Zephyr-ORPO) applied odds ratio preference optimization (Hong et al., 2024b, ORPO) with around 7,000 multi-turn synthetic preference dataset (Daniele & Suphavadeeprasit, 2023). Lambert et al. (2024, TULU3) and OLMo et al. (2025, OLMo2) expanded the general preference alignment into post-instruction-tuning into fine-grained post-training pipeline, especially for complex reasoning.

In our work, we first present a case study on the strategies for constructing synthetic datasets starting from a pre-trained LLM. We go through diverse strategies of teacher distilled supervised fine-tuning (SFT) (Taori et al., 2023; Tunstall et al., 2024): (1) within-model family distillation, (2) mixed-lingual distillation, and (3) multi-turn completions. Consequently, we present our cross-lingual preference optimization pipeline in a scaled model and dataset size, where we do not necessitate large-scale human annotations of a target language. By effectively leveraging the multilingual capabilities of teacher models and performing on-policy preference optimization, we attain state-of-the-art performance bilingual LLM. In this paper, we specifically target Korean as the target language. Our resultant bilingual LLM, *Kowen*, shows comparable results to top-performing Korean LLMs. In contrast to other closed-recipes, we open-source all training data and code for reproducibility.

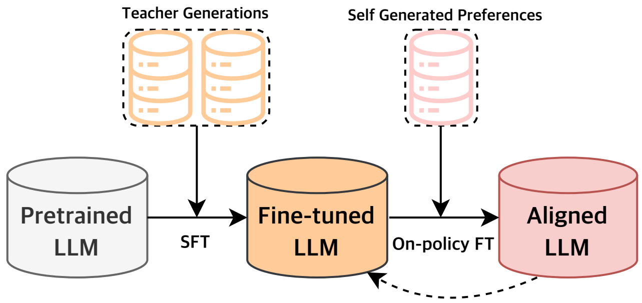


Figure 1: Overall Pipeline of the proposed method. Starting from the pretrained model, we go through supervised fine-tuning (SFT) from the teacher model (Qwen-2.5-72B-Instruct) and iteratively fine-tune the model with self-generated preferences.

2 Harnessing Multilingual Synthetic Data

In this section, we study the impact of important experimental choices in post-training under specific purpose of language specialization in large language models (LLMs).

2.1 Experimental setup

Model and Dataset We select two sizes (7-8B & 70-72B) from two different model families (Qwen2.5 (Qwen et al., 2025) and Llama3.1 (Dubey et al., 2024)) as the teacher model. For the student model, we use the base Qwen2.5-3B.

Given a total budget of 20K training instances, we sample out human-generated prompts from LMSYS-Chat-1M (Zheng et al., 2024). Then, we either generate single-turn completions in English or Korean[1]. To foster multi-turn instruction-following capability in Korean, we adopt a setting similar to that of Xu et al., 2025, where we continually generate the multi-turn query and response based on the previous turn.

Model	Teacher	AE 2.0 (Ko)
Qwen2.5-3B-IT	-	13.0
Qwen2.5-3B	Qwen2.5-7B-IT	19.1
Qwen2.5-3B	Qwen2.5-72B-IT	**21.2**
Qwen2.5-3B	Llama-3.1-8B-IT	4.3
Qwen2.5-3B	Llama-3.1-70B-IT	3.5

Table 1: Korean instruction-following ability evaluation under different teacher sources. We select two different model families, each with two model sizes as teacher models, and select Qwen2.5-3B as a student model.

Training To train the student model, we supervise fine-tune (SFT) on the synthetic generations from the teachers, given an equal dataset budget and training hyperparameters in Appendix A.

Evaluation While we want to build a language model specialized in Korean, we still evaluate its performance on English benchmarks as well to make sure it does not fail in English. For English benchmarks, we evaluate the models on AlpacaEval 2.0 (Dubois et al., 2024) and the Korean translated AlpacaEval (See Appendix B).

2.2 Results

Effect of Teacher Model Family From Table 1, we can see a dramatic synergy of utilizing synthetic generations from the same model family. While utilizing the generations from the Llama3.1 family does not incur notable benefits, generations from Llama-3.1-70B-IT result in

[1]We translate the prompts first in Korean using X-ALMA(Xu et al., 2024)

Models	English			Korean		
	AE 2.0	**MTB** (1^{st})	**MTB** (2^{nd})	**AE 2.0**	**MTB** (1^{st})	**MTB** (2^{nd})
Qwen2.5-7B-IT (Qwen et al., 2025)	30.3	7.5	7.0	38.8	8.2	7.4
Llama-3.1-8B-IT (Dubey et al., 2024)	31.5	8.6	7.7	13.9	6.3	5.8
Gemma-2-9B-IT (Team et al., 2024)	47.5	8.7	8.1	62.4	8.5	8.0
Exaone-3.5-7.8B-IT (Research et al., 2024)	**54.2**	**9.3**	8.5	**85.5**	**8.9**	**8.8**
VARCO-8B-IT[2]	25.9	7.0	7.6	50.7	8.6	8.5
Kowen-7B-IT (Ours)	41.8	**9.3**	**8.7**	75.0	**8.9**	8.6

Table 2: English and Korean instruction-following ability assessment of Kowen and five open-source language models of similar sizes through AlpacaEval 2.0 and Multi-Turn Benchmark.

the worst Korean AlpacaEval win-rate, falling 15.6% short of a 7B model from the Qwen2.5 family. In contrast, utilizing the generations from the same model family incurred a great boost in performance even though the Instruct checkpoint has gone through extensive post-training of SFT, offline and online reinforcement learning (RL). Therefore, simply utilizing the same model family contributes to more effective distillation even though the model sizes and training details differ.

Effect of Language Mixing The results in Table 3 interestingly show how the data mixture does not effect largely on the English AlpacaEval 2.0 performance while it greatly benefits on the Korean benchmark. Fine-tuning the Qwen2.5-3B model on the Korean-only mixture of 20K instances ranked second in the English benchmark while using English-only ranked last in the Korean benchmark. Thus, we conjecture the English-dominant training distribution of the original checkpoint requires more non-English and diverse synthetic data to attain competitive bilingual capability.

Dataset Mixture	English	Korean
	AE 2.0	**AE 2.0**
Ko (ST)	13.3	21.2
En (ST)	**14.5**	19.1
Ko (ST) + En (ST)	12.9	21.5
Ko (ST) + Ko (MT) + En (ST)	13.2	**28.7**

Table 3: English and Korean instruction-following ability evaluations for different multilingual synthetic data setups. "Ko" and "En" refer to Korean and English-only data, and "ST" and "MT" denote single and multi-turn data.

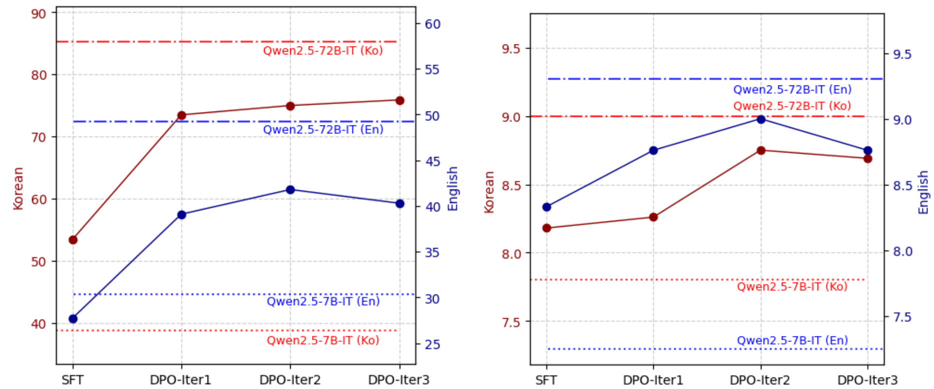
3 Kowen: Training a SOTA Bilingual LLM from a Pre-trained LLM

In this section, we introduce the two-step training pipeline for Kowen, comprising supervised fine-tuning (SFT) as distillation and iterative preference alignment with self-generations following the insights we introduced in Section 2.

3.1 Stage 1. SFT as Distillation

Model We use the Qwen2.5 family (Qwen et al., 2025) to conduct SFT to distill the Korean capabilities. We select Qwen2.5-72B-Instruct as a teacher and the Qwen2.5-7B base as a student model.

Dataset We now use the entire 1M instances from LMSYS-Chat-1M (Zheng et al., 2024). We design chat completions from teacher model that more resemble the real-world use cases: (1) complex single-turn conversations; (2) multi-turn conversations. Our final 1M dataset composes: (1) 400K single-turn Korean responses, (2) 300K single-turn English responses, and (3) 400K multi-turn Korean responses, a mixture based on findings in Section 2.



(a) English & Korean AlpacaEval 2.0 (b) English & Korean MT-Bench

Figure 2: Evaluating instruction-following abilities in Korean and English with AlpacaEval 2.0 (Figure 2a) and MT-Bench (Figure 2b) throughout the iterative alignment process with self-generation. The y-scales at the left (red) and right (blue) represents the Korean and English scales respectively.

3.2 Stage 2. Iterative Alignment with Self-Generations

Adopting similar approaches done in Meng et al., 2024; Hong et al., 2024a, we utilize self-generated responses for fine-tuning. We go through an iterative training process on the Direct Preference Optimization objective (Rafailov et al., 2024, DPO).

Dataset We take the prompts from the cleaned UltraFeedback (Bartolome et al., 2023; Cui et al., 2024, UF) of 60k, and translate them into Korean using GPT-4o. Then, we generate four chat completions from the seed model and generate the reward for each completions using the reward model. Finally, we select the responses with the highest and lowest reward to construct the preference pairs for fine-tuning.

Training Setup Under the DPO objective below:

$$-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right), \tag{1}$$

we iteratively train the SFT model from Section 3.1 to a maximum of three iterations[3] with the UF prompts and corresponding self-generated preference pairs as illustrated in Figure 1.

Evaluation For a more comprehensive evaluation, on top of AlpacaEval, we also evaluate the models on Multi-Turn Benchmark (Zheng et al., 2023, MT-Bench) and its translated version Ko-MTBench (Research, 2024) for evaluation on multi-turn instruction following performance. Also, we evaluate on two natural language understanding benchmarks: MMLU (Hendrycks et al., 2021), KMMLU (Son et al., 2024) and Belebele (English and Korean) (Bandarkar et al., 2023) using the `lm-evaluation-harness` tool (Gao et al., 2024).

3.3 Results

Kowen performs well in both English & Korean Benchmarks As shown in Tables 2 and 4, our resulting model, Kowen, shows strong English and Korean capability both in instruction-following and NLU benchmarks. Especially for AlpacaEval, Kowen achieves 11.5 % and 36.2 % higher than the Qwen2.5-7B-IT model respectively, surprising as they both were trained from the same pre-trained checkpoint. Kowen ranks first in the MT-Bench with the exception of the 2nd turn of Korean MT-Bench. On the other hand, Exaone-3.5-7.8B-IT performs better in instruction-following benchmarks, especially in AlpacaEval

[3]We select the checkpoint trained up to the 2nd iteration as the final checkpoint.

Models	MMLU		Belebele	
	En	Ko	En	Ko
Qwen2.5-7B-IT	0.71	0.48	**0.91**	**0.84**
Exaone-3.5-7.8B-IT	0.65	0.46	0.84	0.71
VARCO-8B-IT	0.62	0.37	0.86	0.75
Kowen-7B-IT (Ours)	**0.72**	**0.50**	0.88	0.81

Table 4: English and Korean natural language understanding (NLU) assessment of Kowen and three open-source language models of similar sizes.

where the main focus is on general instruction-following capability, but not as well in the NLU evaluations. In NLU evaluations, Kowen also ranks first with both the Korean and English MMLU benchmarks. Qwen2.5-7B-IT marginally leads the Belebele benchmark with a margin of 0.03 in both English and Koream.

Iterative DPO in Korean increases English capability as well In Figure 2, we visualize the performance trend of our training pipeline. It can be seen how the initial SFT stage alone yields a model stronger than the Qwen2.5-7B-IT checkpoint with the exception of the result for English AlpacaEval. Through iterations, the performance rises until the second iteration of DPO, while the third iteration brought a marginal performance increase or drop in most cases. We conjecture the diminishing return of our iterative process to be attributed to our fixed prompts from UltraFeedback. We select DPO-Iter-2 as our final checkpoint, **Kowen**, which showed the strongest bilingual capabilities.

4 Discussion

Leveraging synthetic data for multilingual capability Within the ground of multilingual large language models (MLLMs), recent works showed MLLMs can be specialized to specific languages with synthetic data (Kim et al., 2024; Polignano et al., 2024; Research et al., 2024; Devine, 2024). Research et al. (2024, EXAONE-3.5) presented a mostly closed recipe for their language models pre-trained to be specialized for two languages, English and Korean, utilizing synthetic preference data for the preference alignment phase. Devine (2024) analyzed the effectiveness of the preference learning mechanism as a language specification, proposing the importance of data curation. Despite works outlining the potential of multilingual synthetic data. Meanwhile, VARCO was post-trained on top of Llama-3.1-8B (Dubey et al., 2024) with undisclosed Korean data, unknown whether synthetic data has been used. We expect our work to facilitate further research for use-casing multilingual synthetic data generation with increased openness.

5 Conclusion

We introduce **Kowen**, an open-recipe bilingual LLM effectively distilled with synthetic data. We study the two design choices in synthetic distillation with LLMs: (1) student-teacher model family alignment and (2) language composition. We empirically show that comprehensive use of target language *and* English best specializes language models for the target language. Furthermore, iterative preference alignment with self-generated data effectively leads to a boost in downstream performance. **Kowen**, trained entirely in an open manner, achieves 9.0 and 8.75 in English and Korean MT-Bench, outperforming closed bilingual LLMs, demonstrating the strength of language-specific distillation via iterative alignment.

6 Limitation

Our paper actively utilizes synthetic data generations to specialize into certain languages. However, the limitations and possible side-effects of utilizing synthetic data generations are not heavily studied in our work. We leave for future work to expand on the risks of utilizing synthetic data especially on the ground of multilinguality. Furthermore, our work deals only with Korean and English, languages where multilingual large language models excel at. Case studies investigating the effect of synthetic data in low-resource languages will bring inspiring studies for language specialization for more marginal cases.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.

Alvaro Bartolome, Gabriel Martin, and Daniel Vila. Notus. https://github.com/argilla-io/notus, 2023.

Alvaro Bartolome, Jiwoo Hong, Noah Lee, Kashif Rasul, and Lewis Tunstall. Zephyr 141b a39b. https://huggingface.co/HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BOorDpKHiJ.

Luigi Daniele and Suphavadeeprasit. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training. *arXiv preprint arXiv:(coming soon)*, 2023. URL https://huggingface.co/datasets/LDJnr/Capybara.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient fine-tuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Peter Devine. Are you sure? rank them again: Repeated ranking for better preference datasets. In Jonne Sälevä and Abraham Owodunni (eds.), *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pp. 93–105, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.mrl-1.5. URL https://aclanthology.org/2024.mrl-1.5/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. Cross-lingual transfer of reward models in multilingual alignment. *arXiv preprint arXiv:2410.18027*, 2024a.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. Efficient and effective vocabulary expansion towards multilingual large language models, 2024.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL https://arxiv.org/abs/2405.07101.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020. ISBN 9781728199986.

LG AI Research. Komt-bench. https://huggingface.co/datasets/LGAI-EXAONE/KoMT-Bench, 2024.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. Exaone 3.5: Series of large language models for real-world use cases, 2024. URL https://arxiv.org/abs/2412.04862.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*, 2024.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pnk7vMbznK.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

A Training Configurations

Both SFT and DPO were done using Hugging Face TRL library (von Werra et al., 2020) on 4 NVIDIA A100 GPUs with Accelerate (Gugger et al., 2022) and DeepSpeed ZeRO 3 (Rajbhandari et al., 2020), and Paged AdamW optimizer (Loshchilov & Hutter, 2019; Dettmers et al., 2023) with 8-bit precision (Dettmers et al., 2022).

Supervised Fine-tuning For all supervised fine-tuning (SFT) processes, we used a maximum learning rate of $1e-5$ and 10% of warm-up followed by cosine decay. The global batch was set to 128.

On-Policy Preference Optimization We fine-tune our fine-tuned (via SFT) Qwen2.5-7B (Qwen et al., 2025) iteratively with DPO. We use a cosine decaying learning rate scheduler for single epoch training.

DPO configurations We apply $\beta = 0.1$ for the first iteration and apply $\beta = 1.0$ for the iterations after with the learning rate of $5e-7$. The global batch size was set to 128 using gradient accumulation steps of 16 with a per-device batch size of 2.

B MULTILINGUAL ALPACAEVAL Setup

We use the exactly same setup from Hong et al., 2024a, where we utilize the translated prompt instances[4] and compute the language-specific win-rate of the model evaluated by GPT-4o[5] against the reference responses from GPT-4-Turbo[6].

[4] https://huggingface.co/datasets/zhihz0535/X-AlpacaEval
[5] https://platform.openai.com/docs/models/gpt-4o
[6] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4